Tai-hoon Kim
Young-hoon Lee
Byeong-Ho Kang
Dominik Ślęzak (Eds.)

LNCS 6485

Future Generation Information Technology

Second International Conference, FGIT 2010
Jeju Island, Korea, December 2010
Proceedings

Springer

Lecture Notes in Computer Science 6485

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Tai-hoon Kim Young-hoon Lee
Byeong-Ho Kang Dominik Ślęzak (Eds.)

Future Generation Information Technology

Second International Conference, FGIT 2010
Jeju Island, Korea, December 13-15, 2010
Proceedings

Springer

Volume Editors

Tai-hoon Kim
Hannam University, Daejeon, South Korea
E-mail: taihoonn@hnu.kr

Young-hoon Lee
Hannam University, Daejeon, South Korea
E-mail: yhnlee@sogang.ac.kr

Byeong-Ho Kang
University of Tasmania, Hobart, Australia
E-mail: byeong.kang@utas.edu.au

Dominik Ślęzak
University of Warsaw & Infobright, Warsaw, Poland
E-mail: slezak@infobright.com

Library of Congress Control Number: 2010940214

CR Subject Classification (1998): C.2, H.4, I.2, I.4, D.2, H.3

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743
ISBN-10 3-642-17568-6 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-17568-8 Springer Berlin Heidelberg New York

springer.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 06/3180

Preface

As information technology (IT) becomes specialized and fragmented, it is easy to lose sight that many topics have common threads and because of this, advances in one subdiscipline may transmit to another. The presentation of results between different subdisciplines encourages this interchange for the advancement of IT as a whole.

This volume comprises the selection of papers presented at the Second International Mega-Conference on Future Generation Information Technology (FGIT 2010), composed of the following 11 international conferences: Advanced Software Engineering and Its Applications (ASEA 2010), Bio-Science and Bio-Technology (BSBT 2010), Control and Automation (CA 2010), Disaster Recovery and Business Continuity (DRBC 2010), Database Theory and Application (DTA 2010), Future Generation Communication and Networking (FGCN 2010), Grid and Distributed Computing (GDC 2010), Multimedia, Computer Graphics and Broadcasting (MulGraB 2010), Security Technology (SecTech 2010), Signal Processing, Image Processing and Pattern Recognition (SIP 2010), as well as u- and e-Service, Science and Technology (UNESST 2010).

In total, 1,630 papers were submitted to FGIT 2010 from 30 countries. The submitted papers went through a rigorous reviewing process and 395 papers were accepted. Of these 395 papers, 60 were assigned to this volume. In addition, this volume contains 7 invited papers and abstracts. Of the remaining accepted papers, 269 were distributed among 8 volumes of proceedings published by Springer in the CCIS series. 66 papers were withdrawn due to technical reasons.

We would like to acknowledge the great effort of all in the International Advisory Boards and International Program Committees, as well as the organizations and individuals who supported the idea of publishing this volume, including SERSC and Springer. Also, the success of FGIT 2010 would not have been possible without the huge support from our sponsors and the work of the Chairs and Organizing Committee.

We are grateful to the following keynote speakers who kindly accepted our invitation: Hojjat Adeli (Ohio State University), Ruay-Shiung Chang (National Dong Hwa University), and Andrzej Skowron (University of Warsaw). We would also like to thank all plenary and tutorial speakers for their valuable contributions.

We would like to express our greatest gratitude to the authors and reviewers of all paper submissions, as well as to all attendees, for their input and participation.

Last but not least, we give special thanks to Rosslin John Robles and Maricel Balitanas. These graduate school students of Hannam University contributed to the editing process of this volume with great passion.

December 2010

Tai-hoon Kim
Young-hoon Lee
Byeong-Ho Kang
Dominik Ślęzak

Organization

Honorary Chair

Young-hoon Lee	Hannam University, Korea

Steering Co-chairs

Tai-hoon Kim	Hannam University, Korea
Wai-chi Fang	National Chiao Tung University, Taiwan

International Advisory Board

Haeng-kon Kim	Catholic University of Daegu, Korea
Tughrul Arslan	Edinburgh University, UK
Adrian Stoica	NASA Jet Propulsion Laboratory, USA
Yanchun Zhang	Victoria University, Australia
Stephen S. Yau	Arizona State University, USA
Sankar K. Pal	Indian Statistical Institute, India
Jianhua Ma	Hosei University, Japan
Aboul Ella Hassanien	Cairo University, Egypt

Program Chair

Dominik Ślęzak	University of Warsaw and Infobright, Poland

Program Co-chairs

Byeong-Ho Kang	University of Tasmania, Australia
Akingbehin Kiumi	University of Michigan-Dearborn, USA
Xiaofeng Song	Nanjing University of Aeronautics and Astronautics, China
Kyo-il Chung	ETRI, Korea
Kirk P. Arnett	Mississippi State University, USA
Frode Eika Sandnes	Oslo University College, Norway

Publicity Co-chairs

Junzhong Gu	East China Normal University, China
Hideo Kuroda	Nagasaki University, Japan
Dae-sik Ko	Mokwon University, Korea
Minsuk O.	Kyunggi University, Korea
Robert C.H. Hsu	Chung Hua University, Taiwan

Publication Chair

Bongen Gu	Chungju National University, Korea

Table of Contents

Keynote and Plenary Papers

Data Analysis, Data Processing, Advanced Computation Models

Security, Software Engineering, Communication and Networking

Keynote and Plenary Papers (Continued)

Wavelet-Chaos-Neural Network Models for EEG-Based Diagnosis of Neurological Disorders

Hojjat Adeli

Abba G. Lichtenstein Professor, Departments of Biomedical Engineering, Biomedical Informatics, Civil and Environmental Engineering and Geodetic science, Electrical and Computer Engineering, Neurological Surgery, and Neuroscience, The Ohio State University
470 Hitchcock Hall, 2070 Neil Avenue, Columbus, Ohio 43110 U.S.A.

Abstract. In this Keynote Lecture an overview of the author's research for automated electroencephalogram (EEG)-based diagnosis of neurological disorders is presented. Sample research and wavelet-chaos-neural network models developed by the author and his research associates in recent years for diagnosis of epilepsy, Attention Deficit Hyperactivity Disorder (ADHD), Autism Spectrum Disorder (ASD), and the Alzheimer's Disease (AD) are reviewed briefly. The significant impact of this research on the future of neurology practice and its ramification are discussed.

1 Introduction

A focus of the author's research and his associates since 2002 has been development of novel algorithms for automated electroencephalogram (EEG)-based diagnosis of neurological disorders [2, 15]. These are complicated diagnostic problems that are considered too difficult for a machine to solve. It was perceived for a long time that only highly trained neurologists and epileptologists are able to *read* EEGs and make the diagnosis primarily for one specific type of neurological disorder, that is, epilepsy and seizure. EEGs are not commonly used by neurologists for diagnosis of other neurological disorders. This Keynote Lecture presents an overview of the author's odyssey, investigation, and discovery in this area. The general approach adopted by the author is a multi-paradigm methodology for the complicated problems of the EEG-based diagnosis of neurological and psychiatric disorders, through adroit integration of chaos theory based on nonlinear science, wavelets, a signal processing technique [10, 34-35, 44, 86, 100-103], and soft computing techniques such as neural networks [7-9, 11-14, 16, 22-24, 29-32, 42-43, 45-53, 97] and fuzzy logic [25, 54, 80, 85, 87-88, 94, 98]. A parallel line of research is development of more biologically realistic spiking neural networks [36-38].

2 EEG-Based Diagnosis of Epilepsy and Seizure Detection

About 1% of the people in the world suffer from epilepsy and 30% of epileptics are not helped by medication. Careful analyses of the EEG records can provide valuable

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 1–11, 2010.

insight and improved understanding of the mechanisms causing epileptic disorders. Wavelet transform is particularly effective for representing various aspects of non-stationary signals such as trends, discontinuities, and repeated patterns where other signal processing approaches fail or are not as effective. In 2003, Adeli et al. [15] investigated application of discrete Daubechies and harmonic wavelets for analysis of epileptic EEG records. They used wavelet transform to analyze and characterize epileptiform discharges in the form of three-Hz spike and wave complex in patients with absence seizure. Through wavelet decomposition of the EEG records, transient features are accurately captured and localized in both time and frequency context. The capability of this mathematical microscope to analyze different scales of neural rhythms is shown to be a powerful tool for investigating small-scale oscillations of the brain signals. They concluded that wavelet analyses of EEGs obtained from a population of patients can potentially suggest the physiological processes undergoing in the brain in epilepsy onset. Other researchers have followed this line of research in recent years [1, 26, 28, 33, 41, 66-67, 72-73, 81, 93, 99].

Recently Adeli et al. presented a wavelet-chaos methodology for analysis of EEGs and *delta*, *theta*, *alpha*, *beta*, and *gamma* sub-bands of EEGs for detection of seizure and epilepsy [5]. The non-linear dynamics of the original EEGs are quantified in the form of the correlation dimension (CD, representing system complexity) and the largest Lyapunov exponent (LLE, representing system chaoticity). The wavelet-based methodology isolates the changes in CD and LLE in specific sub-bands of the EEG. The methodology is applied to three different groups of EEG signals: (a) healthy subjects (b) epileptic subjects during a seizure-free interval (interictal EEG), and (c) epileptic subjects during a seizure (ictal EEG). The effectiveness of CD and LLE in differentiating between the three groups is investigated based on statistical significance of the differences. It is observed that while there may not be significant differences in the values of the parameters obtained from the original EEG, differences may be identified when the parameters are employed in conjunction with specific EEG sub-bands.

Ghosh-Dastidar et al. [39] present a mixed-band wavelet-chaos-neural network methodology for epilepsy and epileptic seizure detection. Wavelet analysis is used to decompose the EEG into *delta*, *theta*, *alpha*, *beta*, and *gamma* sub-bands. Three parameters are employed for EEG representation: standard deviation (quantifying the signal variance), correlation dimension, and largest Lyapunov exponent. A number of different classifiers were investigated and their classification accuracies were compared. It is concluded that all three key components of the wavelet-chaos-neural network methodology are important for improving the EEG classification accuracy. Judicious combinations of parameters and classifiers are needed to accurately discriminate between the three types of EEGs. It was discovered that a particular mixed-band feature space consisting of nine parameters and Levenberg-Marquardt backpropagation neural network result in the highest classification accuracy, a high value of 96.7%.

Ghosh-Dastidar et al. [40] present a principal component analysis (PCA)-enhanced cosine radial basis function neural network (RBFNN) classifier. The two-stage classifier is integrated with the mixed-band wavelet-chaos methodology, described earlier, for accurate and robust classification of EEGs into healthy, ictal, and interictal EEGs. A nine-parameter mixed-band feature space discovered in previous research for effective EEG representation is used as input to the two-stage classifier. In the first stage, PCA is employed for feature enhancement. The rearrangement of the input space along the principal components of the data improves the classification accuracy of the cosine RBFNN employed in the second stage significantly. The wavelet-chaos-neural network methodology yields high EEG classification accuracy (96.6%) and is robust to changes in training data with a low standard deviation of 1.4%.

3 EEG-Based Diagnosis of Attention Deficit/Hyperactivity Disorder (ADHD)

Ahmadlou and Adeli [18] present a multi-paradigm methodology for EEG-based diagnosis of Attention-Deficit/Hyperactivity Disorder (ADHD) through integration of chaos theory, wavelets [57-58, 60-62, 69, 96], and neural networks [55-56, 63-65, 68, 70-71, 74-78, 82-84, 89-90, 92, 95]. The selected chaotic features are generalized synchronizations known as synchronization likelihoods (SL), both among all electrodes and among electrode pairs. The methodology consists of three parts: first detecting the more synchronized loci (group 1) and loci with more discriminative deficit connections (group 2). Using SLs among all electrodes, discriminative SLs in certain sub-bands are extracted. In part two, SLs are computed, not among all electrodes, but between loci of group 1 and loci of group 2 in all sub-bands and the band-limited EEG. This part leads to more accurate detection of deficit connections, and not just deficit areas, but more discriminative SLs in sub-bands with finer resolutions. In part three, a classification technique, radial basis function neural network [59, 79], is used to distinguish ADHD from normal subjects. The methodology was applied to EEG data obtained from 47 ADHD and 7 control individuals with eyes closed. Using the RBF neural network classifier the methodology yielded a high accuracy of 96.5% for diagnosis of the ADHD.

Ahmadlou and Adeli [17] present a new measure of synchronization, called fuzzy SL, using the theory of fuzzy logic and Gaussian membership functions. The new fuzzy SL is applied to EEG-based diagnosis of ADHD. The results of ANOVA analysis indicate the interdependencies measured by the fuzzy SL are more reliable than the conventional SL for discriminating ADHD patients from healthy individuals.

4 EEG-Based Diagnosis of the Alzheimer's Disease

Prediction or early-stage diagnosis of Alzheimer's disease (AD) requires a comprehensive understanding of the underlying mechanisms of the disease and its progression. Researchers in this area have approached the problem from multiple directions by attempting to develop (a) neurological (neurobiological and neurochemical) models, (b) analytical models for anatomical and functional brain images, (c) analytical feature extraction models for electroencephalograms (EEGs), (d) classification models for positive identification of AD, and (e) neural models of memory and memory impairment in AD [4]. Adeli et al. [4] present a review of research performed on computational modeling of AD and its markers, covering computer imaging, classification models, connectionist neural models, and biophysical neural models. A subject of significant research interest is detecting markers in EEGs obtained from AD patients. Adeli et al. [3] present a review of models of computation and analysis of EEGs for diagnosis and detection of AD. This review covers three areas: time-frequency analysis, wavelet analysis, and chaos analysis.

Adeli et al. [6] present a spatio-temporal wavelet-chaos methodology for analysis of EEGs and their *delta*, *theta*, *alpha*, and *beta* sub-bands for discovering potential markers of abnormality in Alzheimer's disease (AD). The non-linear dynamics of the EEG and EEG sub-bands are quantified in the form of CD and LLE. The methodology is applied to two groups of EEGs: healthy subjects and AD patients. The eyes open and eyes closed conditions are investigated to evaluate the effect of visual input and attention. EEGs from different loci in the brain are investigated to discover areas of the brain responsible for or affected by changes in CD and LLE. It is found that the wavelet-chaos methodology and the sub-band analysis developed in the research accurately characterize the nonlinear dynamics of non-stationary EEG-like signals with respect to the EEG complexity and chaoticity.

Sarkari et al. [91] present a comprehensive EEG study for interhemispheric, intrahemispheric and distal coherence in AD patients. EEGs are obtained from 20 AD probable patients and 7 healthy (control) subjects. Pair-wise electrode coherence is calculated over each frequency band (delta, theta, alpha and beta). One-way Analysis of Variation (ANOVA) test shows a set of statistically significant differences in electrode coherence between AD and controls.

Ahmadlou et al. [19] use fractal dimension (FD) for evaluation of the dynamical changes in the AD brain. First, two different FD algorithms for computing the fractality of EEGs are investigated and their efficacy for yielding potential mathematical markers of AD is compared. They are Katz's FD (KFD) and Higuchi's FD (HFD). Significant features in different loci and different EEG sub-bands or band-limited EEG for discrimination of AD and control group are determined by ANOVA test. The most discriminative FD and the corresponding loci and EEG sub-bands for discriminating between AD and healthy EEGs are discovered.

Ahmadlou et al. [20] present a chaos-wavelet approach for EEG-based diagnosis of AD employing a recently developed concept in graph theory, Visibility Graph (VG). Complexity of EEGs is computed using the VGs of EEGs and EEG sub-bands produced by wavelet decomposition. Two methods are employed for computation of complexity of the VGs: one based on the power of scale-freeness of a graph structure and the other based on the maximum eigenvalue of the adjacency matrix of a graph. Two classifiers are applied to the selected features to distinguish AD and control EEGs: a Radial Basis Function Neural Network (RBFNN) and a two-stage classifier consisting of Principal Component Analysis (PCA) and the RBFNN. After comprehensive statistical studies, effective classification features and mathematical markers were discovered.

5 EEG-Based Diagnosis of the Autism Spectrum Disorder

Ahmadlou et al. [21] present a method for investigation of EEGs of children with autistic spectrum disorder (ASD) using complexity and chaos theory with the goal of discovering a nonlinear feature space. Fractal Dimension (FD) is proposed for investigation of complexity and dynamical changes in ASD brain. Two methods are investigated for computation of FD: HFD and KFD described earlier. The model is tested on a database of eyes-closed EEG data obtained from two groups: 9 ASD and 8 non-ASD children. Using a radial basis function classifier, an accuracy of 90% was achieved based on the most significant features discovered via ANOVA statistical test.

Acknowledgement

The following research associates and collaborators have contributed to research presented in this Keynote Lecture as published in a recent book and a number of journal articles, and cited in this overview article. Mr. Mehran Ahmadlou, Dr. Nahid Dadmehr, Dr. Sam Ghosh-Dastidar, Mr. Ziad Sankari, and Ms. Ziqin Zhou.

References

1. Acharya, R., Chua, E.C.P., Chua, K.C., Min, L.C., Tamura, T.: Analysis and Automatic Identification of Sleep Stages using Higher Order Spectra. International Journal of Neural Systems 20(6) (2010)
2. Adeli, H., Ghosh-Dastidar, S.: Automated EEG-based Diagnosis of Neurological Disorders - Inventing the Future of Neurology. CRC Press, Taylor & Francis, Boca Raton, Florida (2010)
3. Adeli, H., Ghosh-Dastidar, S., Dadmehr, N.: Alzheimer's Disease: Models of Computation and Analysis of EEGs. Clinical EEG and Neuroscience 36(3), 131–140 (2005)
4. Adeli, H., Ghosh-Dastidar, S., Dadmehr, N.: Alzheimer's Disease and Models of Computation: Imaging, Classification, and Neural Models. Journal of Alzheimer's Disease 7(3), 187–199 (2005)
5. Adeli, H., Ghosh-Dastidar, S., Dadmehr, N.: A Wavelet-Chaos Methodology for Analysis of EEGs and EEG Sub-bands to detect Seizure and Epilepsy. IEEE Transactions on Biomedical Engineering 54(2), 205–211 (2007)
6. Adeli, H., Ghosh-Dastidar, S., Dadmehr, N.: A Spatio-temporal Wavelet-Chaos Methodology for EEG-based Diagnosis of Alzheimer's Disease. Neuroscience Letters 444(2), 190–194 (2008)
7. Adeli, H., Hung, S.L.: An Adaptive Conjugate Gradient Learning Algorithm for Effective Training of Multilayer Neural Networks. Applied Mathematics and Computation 62(1), 81–102 (1994)
8. Adeli, H., Karim, A.: Scheduling/Cost Optimization and Neural Dynamics Model for Construction. Journal of Construction Management and Engineering, ASCE 123(4), 450–458 (1997)
9. Adeli, H., Karim, A.: Fuzzy-Wavelet RBFNN Model for Freeway Incident Detection. Journal of Transportation Engineering 126(6), 464–471 (2000)
10. Adeli, H., Kim, H.: Wavelet-Hybrid Feedback Least Mean Square Algorithm for Robust Control of Structures. Journal of Structural Engineering, ASCE 130(1), 128–137 (2004)
11. Adeli, H., Park, H.S.: Counter Propagation Neural Network in Structural Engineering. Journal of Structural Engineering, ASCE 121(8), 1205–1212 (1995)
12. Adeli, H., Park, H.S.: A Neural Dynamics Model for Structural Optimization - Theory. Computers and Structures 57(3), 383–390 (1995)

13. Adeli, H., Park, H.S.: Optimization of Space Structures by Neural Dynamics. Neural Networks 8(5), 769–781 (1995)
14. Adeli, H., Samant, A.: An Adaptive Conjugate Gradient Neural Network - Wavelet Model for Traffic Incident Detection. Computer-Aided Civil and Infrastructure Engineering 15(4), 251–260 (2000)
15. Adeli, H., Zhou, Z., Dadmehr, N.: Analysis of EEG Records in an Epileptic Patient Using Wavelet Transform. Journal of Neuroscience Methods 123(1), 69–87 (2003)
16. Ahmadlou, M., Adeli, H.: Enhanced Probabilistic Neural Network with Local Decision Circles: A Robust Classifier. Integrated Computer-Aided Engineering 17(3), 197–210 (2010)
17. Ahmadlou, M., Adeli, H.: Fuzzy Synchronization Likelihood with Application to Attention-Deficit/Hyperactivity Disorder. Clinical EEG and Neuroscience 42(1) (2011)
18. Ahmadlou, M., Adeli, H.: Wavelet-synchronization methodology: a new approach for EEG-based diagnosis of ADHD. Clinical EEG and Neuroscience 41(1), 1–10 (2010)
19. Ahmadlou, A., Adeli, H., Adeli, A.: Fractality and a Wavelet-Chao Methodology for EEG-based Diagnosis of Alzheimer's Disease. Alzheimer Disease and Associated Disorders 24(4) (October 2010)
20. Ahmadlou, M., Adeli, H., Adeli, A.: New Diagnostic EEG Markers of the Alzheimer's Disease Using Visibility Graph. Journal of Neural Transmission 117(9), 1099–1109 (2010)
21. Ahmadlou, M., Adeli, H., Adeli, A.: Fractality and a Wavelet-Chaos-Neural Network Methodology for EEG-based Diagnosis of Autistic Spectrum Disorder. Journal of Clinical Neurophysiology 27(5), 328–333 (2010)
22. Banchs, R., Klie, H., Rodriguez, A., Thomas, S.G., Wheeler, M.F.: A Neural Stochastic Multiscale Optimization Framework for Sensor-based Parameter Estimation. Integrated Computer-Aided Engineering 14(3), 213–223 (2007)
23. Benedic, Y., Wira, P., Merckle, J.: A New Method for the Re-Implementation of Threshold Logic Functions with Cellular Neural Networks. International Journal of Neural Systems 18(4), 293–303 (2008)
24. Boto-Giralda, D., Díaz-Pernas, F.J., González-Ortega, D., Díez-Higuera, J.F., Antón-Rodríguez, M., Martínez-Zarzuela, M.: Wavelet-Based Denoising for Traffic Volume Time Series Forecasting with Self-Organizing Neural Networks. Computer-Aided Civil and Infrastructure Engineering 25(7) (2010)
25. Carden, E.P., Brownjohn, J.M.W.: Fuzzy Clustering of Stability Diagrams for Vibration-Based Structural Health Monitoring. Computer-Aided Civil and Infrastructure Engineering 23(5), 360–372 (2008)
26. Chakravarthy, N., Sabesan, S., Tsakalis, K., Iasemidis, L.: Controling Synchronization in a Neural-level Population Model. International Journal of Neural Systems 17(2), 123–138 (2007)
27. Chen, M., Jiang, C.S., Wu, Q.X., Chen, W.H.: Synchronization in Arrays of Uncertain Delay Neural Networks by Decentralized Feedback Control. International Journal of Neural Systems 17(2), 115–122 (2007)
28. Chiappalone, M., Vato, A., Berdondini, L., Koudelka, M., Martinoia, S.: Network Dynamics and Synchronous Activity in Cultured Cortical Neurons, vol. 17(2), pp. 87–103 (2007)
29. Christodoulou, M.A., Kontogeorgou, C.: Collision Avoidance in commercial aircraft free fight, via neural networks and non-linear programming. International Journal of Neural Systems 18(5), 371–387 (2008)

30. Dharia, A., Adeli, H.: Neural Network Model for Rapid Forecasting of Freeway Link Travel Time. Engineering Applications of Artificial Intelligence 16(7-8), 607–613 (2003)
31. Elragal, H.M.: Improving Neural Networks Prediction accuracy Using Particle Swarm Optimization Combiner. International Journal of Neural Systems 19(5), 387–393 (2009)
32. Fatehi, A., Abe, K.: Flexible Structure Multiple Modeling Using Irregular Self-Organizing Maps Neural Network. International Journal of Neural Systems 18(3), 233–256 (2008)
33. Faust, O., Acharya, U.R., Min, L.C., Sputh, B.H.C.: Automatic Identification of Epileptic and Background EEG Signals Using Frequency Domain Parameters. International Journal of Neural Systems 20(2), 159–176 (2010)
34. Ghosh, B., Basu, B., O'Mahony, M.: Random Process Model for Traffic Flow Using a Wavelet - Bayesian Hierarchical Technique. Computer-Aided Civil and Infrastructure Engineering 25(8) (2010)
35. Ghosh-Dastidar, S., Adeli, H.: Wavelet-Clustering-Neural Network Model for Freeway Incident Detection. Computer-Aided Civil and Infrastructure Engineering 18(5), 325–338 (2003)
36. Ghosh-Dastidar, S., Adeli, H.: Improved Spiking Neural Networks for EEG Classification and Epilepsy and Seizure Detection. Integrated Computer-Aided Engineering 14(3), 187–212 (2007)
37. Ghosh-Dastidar, S., Adeli, H.: Spiking Neural Networks. International Journal of Neural Systems 19(4), 295–308 (2009)
38. Ghosh-Dastidar, S., Adeli, H.: A New Supervised Learning Algorithm for Multiple Spiking Neural Networks with Application in Epilepsy and Seizure Detection. Neural Networks 22, 1419–1431 (2009)
39. Ghosh-Dastidar, S., Adeli, H., Dadmehr, N.: Mixed-band Wavelet-Chaos-Neural Network Methodology for Epilepsy and Epileptic Seizure Detection. IEEE Transactions on Biomedical Engineering 54(9), 1545–1551 (2007)
40. Ghosh-Dastidar, S., Adeli, H., Dadmehr, N.: Principal Component Analysis-Enhanced Cosine Radial Basis Function Neural Network for Robust Epilepsy and Seizure Detection. IEEE Transactions on Biomedical Engineering 55(2), 512–518 (2008)
41. Good, L.B., Sabesan, S., Marsh, S.T., Tsakalis, K.S., Iasemidis, L.D.: Control of Synchronization of Brain Dynamics Leads to Control of Epileptic Seizures in Rodents. International Journal of Neural Systems 19(3), 173–196 (2009)
42. Graf, W., Freitag, S., Kaliske, M., Sickert, J.U.: Recurrent neural networks for uncertain time-dependent structural behavior. Computer-Aided Civil and Infrastructure Engineering 25(5), 322–333 (2010)
43. Haidar, A.M.A., Mohamed, A., Al-Dabbagh, M., Aini Hussain, A., Masoum, M.: An Intelligent Load Shedding Scheme Using Neural networks & Neuro-Fuzzy. International Journal of Neural Systems 19(6), 473–479 (2009)
44. He, Z., You, X., Zhou, L., Cheung, Y., Tang, Y.Y.: Writer Identification Using Fractal Dimension of Wavelet Subbands in Gabor Domain. Integrated Computer-Aided Engineering 17(2), 157–165 (2010)
45. Hooshdar, S., Adeli, H.: Toward Intelligent Variable Message Signs in Freeway Work Zones: A Neural Network Approach. Journal of Transportation Engineering 130(1), 83–93 (2004)
46. Hung, S.L., Adeli, H.: Parallel Backpropagation Learning Algorithms on Cray Y-MP8/864 Supercomputer. Neurocomputing 5(6), 287–302 (1993)
47. Hung, S.L., Adeli, H.: A Parallel Genetic/Neural Network Learning Algorithm for MIMD Shared Memory Machines. IEEE Transactions on Neural Networks 5(6), 900–909 (1994)

48. Hung, S.L., Adeli, H.: Object-Oriented Back Propagation and Its Application to Structural Design. Neurocomputing 6(1), 45–55 (1994)
49. Iglesias, J., Villa, A.E.P.: Emergence of Preferred Firing Sequences in Large Spiking Neural Networks During Simulated Neuronal Development. International Journal of Neural Systems 18(4), 267–277 (2008)
50. Isokawa, T., Nishimura, H., Kamiura, N., Matsui, N.: Associative memory in quaternionic hopfield neural network. International Journal of Neural Systems 18(2), 135–145 (2008)
51. Jiang, X., Adeli, H.: Dynamic Wavelet Neural Network for Nonlinear Identification of Highrise Buildings. Computer-Aided Civil and Infrastructure Engineering 20(5), 316–330 (2005)
52. Jiang, X., Adeli, H.: Pseudospectra, MUSIC, and Dynamic Wavelet Neural Network for Damage Detection of Highrise Buildings. International Journal for Numerical Methods in Engineering 71(5), 606–629 (2007)
53. Jiang, X., Adeli, H.: Dynamic Fuzzy Wavelet Neuroemulator for Nonlinear Control of Irregular Highrise Building Structures. International Journal for Numerical Methods in Engineering 74(7), 1045–1066 (2008)
54. Jin, X.L., Doloi, H.: Modelling Risk Allocation Decision-Making in PPP Projects Using Fuzzy Logic. Computer-Aided Civil and Infrastructure Engineering 24(7), 509–524 (2009)
55. Jorgensen, T.D., Haynes, B.P., Norlund, C.C.F.: Pruning Artificial Neural Networks using neural complexity measures. International Journal of Neural Systems 18(5), 389–403 (2008)
56. Karim, A., Adeli, H.: Comparison of the Fuzzy – Wavelet RBFNN Freeway Incident Detection Model with the California Algorithm. Journal of Transportation Engineering 128(1), 21–30 (2002)
57. Karim, A., Adeli, H.: Incident Detection Algorithm Using Wavelet Energy Representation of Traffic Patterns. Journal of Transportation Engineering, ASCE 128(3), 232–242 (2002)
58. Karim, A., Adeli, H.: Fast Automatic Incident Detection on Urban and Rural Freeways Using the Wavelet Energy Algorithm. Journal of Transportation Engineering, ASCE 129(1), 57–68 (2003)
59. Karim, A., Adeli, H.: Radial Basis Function Neural Network for Work Zone Capacity and Queue Estimation. Journal of Transportation Engineering 129(5), 494–503 (2003)
60. Kim, H., Adeli, H.: Hybrid Control of Smart Structures Using a Novel Wavelet-Based Algorithm. Computer-Aided Civil and Infrastructure Engineering 20(1), 7–22 (2005)
61. Kim, H., Adeli, H.: Wavelet Hybrid Feedback-LMS Algorithm for Robust Control of Cable-Stayed Bridges. Journal of Bridge Engineering, ASCE 10(2), 116–123 (2005)
62. Kim, H., Adeli, H.: Hybrid Control of Irregular Steel Highrise Building Structures Under Seismic Excitations. International Journal for Numerical Methods in Engineering 63(12), 1757–1774 (2005)
63. Li, Y., Zhang, T.: Global exponential stability of fuzzy interval delayed neural networks with impulses on time scales. International Journal of Neural Systems 19(6), 449–456 (2009)
64. Liang, J., Wang, Z., Liu, X.: Global Synchronization in an Array of Discrete-Time Neural Networks with Mixed Coupling and Time-Varying Delays. International Journal of Neural Systems 19(1), 57–63 (2009)
65. Khashman, A.: Blood Cell Identification Using a Simple Neural Network. International Journal of Neural Systems 18(5), 453–458 (2008)

66. Kramer, M.A., Chang, F.L., Cohen, M.E., Hudson, D., Szeri, A.J.: Synchronization Measures of the Scalp EEG Can Discriminate Healthy from Alzheimers Subjects. International Journal of Neural Systems 17(2), 61–69 (2007)
67. Lee, H., Cichocki, A., Choi, S.: Nonnegative Matrix Factorization for Motor Imagery EEG Classification. International Journal of Neural Systems 17(4), 305–317 (2007)
68. Li, T., Sun, C., Zhao, X., Lin, C.: LMI-Based Asymptotic Stability Analysis of Neural Networks with Time-Varying Delays. International Journal of Neural Systems 18(3), 257–265 (2008)
69. Montejo, L.A., Kowalsky, M.J.: Estimation of Frequency Dependent Strong Motion Duration via Wavelets and its Influence on Nonlinear Seismic Response. Computer-Aided Civil and Infrastructure Engineering 23(4), 253–264 (2008)
70. Montina, A., Mendoza, C., Arecchi, F.T.: Role of Refractory Period in Homoclinic Models of Neural Synchronization. International Journal of Neural Systems 17(2), 79–86 (2007)
71. Nemissi, M., Seridi, H., Akdag, H.: The Labeled Systems of Multiple Neural Networks. International Journal of Neural Systems 18(4), 321–330 (2008)
72. Osorio, I., Frei, M.G.: Seizure Abatement with Single DC Pulses: Is Phase Resetting at Play? International Journal of Neural Systems 19(3), 149–156 (2009)
73. Osterhage, H., Mormann, F., Wagner, T., Lehnertz, K.: Measuring the Directionality of Coupling: Phase Versus State Space Dynamics and Application to EEG Time Series. International Journal of Neural Systems 17(3), 139–148 (2007)
74. Panakkat, A., Adeli, H.: Neural Network Models for Earthquake Magnitude Prediction using Multiple Seismicity Indicators. International Journal of Neural Systems 17(1), 13–33 (2007)
75. Panakkat, A., Adeli, H.: Recurrent Neural Network for Approximate Earthquake Time and Location Prediction Using Multiple Seismicity Indicators. Computer-Aided Civil and Infrastructure Engineering 24(4), 280–292 (2009)
76. Pande, A., Abdel-Aty, M.: A Computing Approach Using Probabilistic Neural Networks for Instantaneous Appraisal of Rear-End Crash Risk. Computer-Aided Civil and Infrastructure Engineering 23(7), 549–559 (2008)
77. Park, H.S., Adeli, H.: A Neural Dynamics Model for Structural Optimization - Application to Plastic Design of Structures. Computers and Structures 57(3), 391–399 (1995)
78. Park, H.S., Adeli, H.: Distributed Neural Dynamics Algorithms for Optimization of Large Steel Structures. Journal of Structural Engineering, ASCE 123(7), 880–888 (1997)
79. Pedrycz, W., Rai, R., Zurada, J.: Experience-Consistent Modeling for Radial Basis Function Neural Networks. International Journal of Neural Systems 18(4), 279–292 (2008)
80. Perusich, K.: Using Fuzzy Cognitive Maps to Identify Multiple Causes in Troubleshooting Systems. Integrated Computer-Aided Engineering 15(2), 197–206 (2008)
81. Postnov, D.E., Ryazanova, L.S., Zhirin, R.A., Mosekilde, E., Sosnovtseva, O.V.: Noise Controlled Synchronization in Potassium Coupled Neural Networks. International Journal of Neural Systems 17(2), 105–113 (2007)
82. Puscasu, G., Codres, B.: Nonlinear System Identification Based on Internal Recurrent Neural Networks. International Journal of Neural Systems 19(2), 115–125 (2009)
83. Rao, V.S.H., Murthy, G.R.: Global Dynamics of a Class of Complex Valued Neural Networks. International Journal of Neural Systems 18(2), 165–171 (2008)
84. Reuter, U., Moeller, B.: Artificial Neural Networks for Forecasting of Fuzzy Time Series. Computer-Aided Civil and Infrastructure Engineering 25(5), 363–374 (2010)

85. Rigatos, G.G.: Adaptive fuzzy control with output feedback for H-infinity tracking of SISI nonlinear systems. International Journal of Neural Systems 18(4), 305–320 (2008)
86. Rizzi, M., D'Aloia, M., Castagnolo, B.: Computer Aided Detection of Microcalcifications in Digital Mammograms Adopting a Wavelet Decomposition. Integrated Computer-Aided Engineering 16(2), 91–103 (2009)
87. Rokni, S., Fayek, A.R.: A Multi-Criteria Optimization Framework for Industrial Shop Scheduling Using Fuzzy Set Theory. Integrated Computer-Aided Engineering 17(3), 175–196 (2010)
88. Sabourin, C., Madani, K., Bruneau, O.: Autonomous Biped Gait Pattern based on Fuzzy-CMAC Neural Networks. Integrated Computer-Aided Engineering 14(2), 173–186 (2007)
89. Samant, A., Adeli, H.: Feature Extraction for Traffic Incident Detection using Wavelet Transform and Linear Discriminant Analysis. Computer-Aided Civil and Infrastructure Engineering 13(4), 241–250 (2000)
90. Samant, A., Adeli, H.: Enhancing Neural Network Incident Detection Algorithms using Wavelets. Computer-Aided Civil and Infrastructure Engineering 16(4), 239–245 (2001)
91. Sankari, Z., Adeli, H., Adeli, A.: Intrahemispheric, Interhemispheric and Distal EEG Coherence in Alzheimer's Disease. Clinical Neurophysiology (2011), doi:10.1016/j.clinph.2010.09.008 (accepted for publication)
92. Senouci, A.B., Adeli, H.: Resource Scheduling using Neural Dynamics Model of Adeli and Park. Journal of Construction Engineering and Management, ASCE 127(1), 28–34 (2001)
93. Shoeb, A., Guttag, J., Pang, T., Schachter, S.: Non-invasive Computerized System for Automatically Initiating Vagus Nerve Stimulation Following Patient-Specific Detection of Seizures or Epileptiform Discharges. International Journal of Neural Systems 19(3), 157–172 (2009)
94. Stathopoulos, A., Dimitriou, L., Tsekeris, T.: Fuzzy Modeling Approach for Combined Forecasting of Urban Traffic Flow. Computer-Aided Civil and Infrastructure Engineering 23(7), 521–535 (2008)
95. Tang, Y., Miao, Q., Fang, J.A.: Synchronization of stochastic delayed chaotic neural networks with Markovian switching and application in communication. International Journal of Neural Systems 19(1), 43–56 (2009)
96. Umesha, P.K., Ravichandran, R., Sivasubramanian, K.: Crack detection and quantification in beams using wavelets. Computer-Aided Civil and Infrastructure Engineering 24(8), 593–607 (2009)
97. Villaverde, I., Grana, M., d'Anjou, A.: Morphological Neural Networks and Vision Based Simultaneous Localization and Mapping. Integrated Computer-Aided Engineering 14(4), 355–363 (2007)
98. Villar, J.R., de la Cal, E., Sedano, J.: A Fuzzy Logic Based Efficient Energy Saving Approach for Domestic Heating Systems. Integrated Computer-Aided Engineering 16(2), 151–163 (2009)
99. Wang, X., Chen, Y., Bressler, S., Ding, M.: Evaluating Causal Relations Among Multiple Neurobiological Time Series: Blockwise Versus Pairwise Granger Causality. International Journal of Neural Systems 17(2), 71–78 (2007)
100. Yazdani, A., Takada, T.: Wavelet-Based Generation of Energy and Spectrum Compatible Earthquake Time-Histories. Computer-Aided Civil and Infrastructure Engineering 24(8), 623–630 (2009)

101. Zhou, Z., Adeli, H.: Time-frequency signal analysis of earthquake records using Mexican hat wavelets. Computer-Aided Civil and Infrastructure Engineering 18(5), 379–389 (2003)
102. Zhou, Z., Adeli, H.: Wavelet Energy Spectrum for Time-Frequency Localization of Earthquake Energy. International Journal of Imaging Systems and Technology 13(2), 133–140 (2003)
103. Zou, W., Chi, Z., Lo, K.C.: Improvement of Image Classification Using Wavelet Coefficients with Structured-based Neural Network. International Journal of Neural Systems 18(3), 195–205 (2008)

An Introduction to Perception Based Computing

Andrzej Skowron and Piotr Wasilewski

Institute of Mathematics
Warsaw University
Banacha 2, 02-097 Warsaw, Poland
skowron@mimuw.edu.pl

To Professor Lotfi A. Zadeh on His 90th Birthday

Abstract. We discuss basic notions of Perception Based Computing (PBC). Perception is characterized by sensory measurements and ability to apply them to reason about satisfiability of complex vague concepts used, e.g., as guards for actions or invariants to be preserved by agents. Such reasoning is often referred as adaptive judgment. Vague concepts can be approximated on the basis of sensory attributes rather than defined exactly. Approximations usually need to be induced by using hierarchical modeling. Computations require interactions between granules of different complexity, such as elementary sensory granules, granules representing components of agent states, or complex granules representing classifiers that approximate concepts. We base our approach to interactive computations on generalized information systems and rough sets. We show that such systems can be used for modeling advanced forms of interactions in hierarchical modeling. Unfortunately, discovery of structures for hierarchical modeling is still a challenge. On the other hand, it is often possible to acquire or approximate them from domain knowledge. Given appropriate hierarchical structures, it becomes feasible to perform adaptive judgment, starting from sensory measurements and ending with conclusions about satisfiability degrees of vague target guards. Thus, our main claim is that PBC should enable users (experts, researchers, students) to submit domain knowledge, by means of a dialog. It should be also possible to submit hypotheses about domain knowledge to be checked semi-automatically. PBC should be designed more like laboratories helping users in their research rather than fully automatic data mining or knowledge discovery toolkit. In particular, further progress in understanding visual perception – as a special area of PBC – will be possible, if it becomes more open for cooperation with experts from neuroscience, psychology or cognitive science. In general, we believe that PBC will soon become necessity in many research areas.

Keywords: Rough sets, granular computing, interactive computations, perception based computing, information systems, perception attributes, sensory attributes, action attributes, approximation of complex concepts, ontology approximation.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 12–25, 2010.

1 Introduction

In this paper, we discuss some basic issues of Perception Based Computing (PBC). The issue of perception was studied by many famous researchers following approaches such as structuralism, Gestaltism, ecological optics, or constructivism. Our approach is closer to the one proposed by Professor Lotfi A. Zadeh in Computational Intelligence (see, e.g., [28,29,30,31,32,11]). However, because at the moment we do not see a possibility to define fully semantics of the precisiated language proposed by Lotfi Zadeh we propose another approach. Our approach is more application oriented and is using concept ontologies expressed in natural language or small fragments of natural language (which are a bit beyond of ontologies) specific and relevant for considered application problem. In our approach, rules for reasoning under uncertainty should be adaptively discovered from data and domain knowledge. We consider action-oriented perception [1] driven by actions. Goals of initiated action help with selection of an appropriate perceptual interpretation among many ones attached to given information provided by senses/sensors. We also emphasize the role of interactive computations in PBC. Perception is a specific form of interaction of an agent with its environment (see e.g. [1,23]). Perceiving results of conducted actions is an essential part of feedback mechanism and makes adaptive change of a course of actions possible. The considered approach is wider than studied in the visual perception domain [10]. For example, by analogy to visual perception one can attempt to construct softbots acting in Web on the basis of perception characterized by their sensory measurements together with ability to perform reasoning leading from these measurements to conclusions about satisfiability of complex vague concepts used as guards for actions. The guards can be approximated on the basis of measurements performed by sensory attributes only rather than defined exactly. Satisfiability degrees for guards are results of reasoning called as the adaptive judgment. The approximations are induced using hierarchical modeling. Note that, e.g., injecting domain knowledge into a relational database engine [24] will also require approximate modeling of situation perception by domain experts or system users.

Our approach to PBC is primarily based on generalized information systems, rough sets, and granular computing [13,14,16]. Information systems are treated as dynamic granules used for representing results of interaction of attributes with the environment. Two kinds of attributes are distinguished, namely the perception attributes (including sensory attributes) and the action attributes. Sensory attributes are the basic perception attributes, other perception attributes are constructed on the basis of sensory ones. Actions are activated when their guards (being often complex and vague concepts) hold to a satisfactory degree. We show that information systems can be used for modeling more advanced forms of dynamic interactions in hierarchical modeling. The role of hierarchical interactions is emphasized in modeling of interactive computations. Discovery of structure for hierarchical modeling is still a challenge for visual perception, brain informatics or PBC [17]. However, in some applications it is possible to interactively acquire domain knowledge, e.g., in the form of ontology or a simplified description of the structure (see, e.g., [4,24]). Then, by developing approximation methods, the ontology becomes understandable by the system to a satisfactory degree [4,18]. This means that it becomes feasible to use approximated ontology in order to perform the adaptive judgment reasoning described above.

The idea of interactive computing stems from many fields in computer science such as concurrent processes, non-terminating reactive processes (e.g. operating systems), distributed systems, distributed nets and object-oriented programming. It is still in a developing stage and its foundations are not clarified yet. There are at least two main schools of thought, one pioneered by Peter Wegner and another by Yuri Gurevich [7]. Both schools use the notion of *an algorithm* but with a different approach. Wegner's school uses it in the classical Turing's sense, excluding interactive systems from the scope of algorithms and introducing persistent Turing machines (PTMs) for formal description of interactive systems. Gurevich's school expands meaning of algorithms, covering interactive systems and classical algorithms. However, Gurevich claims that the difference is based solely on terminology. For formal descriptions of algorithms, Gurevich introduced abstract state machines (ASMs). ASMs are more powerful than PTMs as they are capable of simulating PTMs, while the opposite does not hold. In addition to strings or matrices, ASMs compute with non-constructive inputs as relational structures (finite graphs). PTMs can only compute with constructive inputs as strings (or matrices written as strings). There is still no consensus between theoreticians on the statement that interactive systems are more powerful than classical algorithms and cannot be simulated by Turing machines. However, the idea of interactive computing seems to be appealing from a practical point of view: interaction with or harnessing the external environment is inevitable to capture (and steer) behavior of systems acting in the real world. For unpredictive and uncontrolled environments it is impossible to specify the exact set of input states. In data mining or machine learning, the most common case is when we start searching for patterns or constructing concepts on the basis of sample of objects since the whole universe of objects (data) is not known or it would be impractical to begin with the basis of the whole object universe.

Interactive systems have huge learning potential and are highly adaptive. Interactive agents adapt dynamically and harness their environment in achieving goals. Interacting algorithms can not only learn knowledge from experience (which is also done by classical non-interacting learning algorithms), they can change themselves during the learning process in response to experience. This property creates an open space for a new technology called Wisdom technology (Wistech) [8]. It becomes inevitable for the case of intelligent agents, which make decisions during dynamic interactions within their environment. To meet this challenge they need to use complex vague concepts. In Wistech, wisdom is a property of algorithms, it is an adaptive ability of making correct judgments to a satisfactory degree in the face of real-life constraints (e.g. time constraints) [8]. These decisions are made on the basis of knowledge possessed by an agent. In Wistech, wisdom is expressed metaphorically by so called *wisdom equation*:

$$\textit{wisdom} = \textit{knowledge} + \textit{adaptive judgment} + \textit{interactions}.$$

Adaptive ability means the ability to improve the judgment process quality taking into account agent experience. Adaptation to the environment on the basis of perceived results of interactions and agent knowledge is needed since, e.g., agents make decisions using concepts which are approximated by classification algorithms (classifiers) and these approximations are changed over time as a result of acting on evolving data and knowledge. The wisdom equation suggests also another interaction of higher order: agents making decisions based on ongoing experience, which is particular, apply possessed knowledge, which is general. Therefore, making decisions itself is a kind of logical interaction between general knowledge and particular experience. Vague concepts in this case help in covering the gap between generality and particularity while Wisdom technology is required to improve decision making.

From the point of view of Wistech PBC, systems should be also interactive with users and domain experts. They should allow users (experts, researchers, or students) for interactions not only for acquiring ontology but also for submitting hypotheses (e.g., about importance or interestingness of some patterns suggested by experts) to be checked by the system. These systems will be more like laboratories helping different experts to make progress in their research rather than performing fully automatic data mining or discovery. Research on visual perception supported by such systems open for cooperation with experts from neuroscience, psychology or cognitive science will become standard. This view is consistent with [5]; See, page 3 of Foreword:

> *Tomorrow, I believe, every biologist will use computer to define their research strategy and specific aims, manage their experiments, collect their results, interpret their data, incorporate the findings of others, disseminate their observations, and extend their experimental observations – through exploratory discovery and modeling – in directions completely unanticipated.*

In our approach, we further combine the-previously mentioned methods based on information systems, rough sets, and granular computing with other soft computing paradigms, such as fuzzy sets or evolutionary computing, as well as data mining and machine learning techniques. Soft computing methods, in particular rough set based methods, are necessary for adaptive approximation of complex vague concepts representing results of agent perception. They are used by agents as, e.g., action guards or invariants which should be preserved by agents. They may be related to highly structural spatio-temporal objects, e.g., functions representing changes of agent states or properties of dynamic processes [19,23].

Information systems play a special role in modeling of interactive computations based on objects called as granules [16]. Granules can be of complex types starting from elementary granules such as indiscernibility or similarity classes [15] to more complex ones such as decision rules, sets of decision rules, classifiers, clusters, time windows or their clusters, sequences of time windows or processes, agents or teams of agents.

In Section 2, we present a general scheme of interactions. The role of attributes and information systems in hierarchical modeling is outlined in Section 3. On the basis of the general scheme of interactions we introduce interactive attributes. In the definition of attribute, two components are important. The first one is defined by a relational structure and the second one is representing a partial information about the results of interactions of the relational structure with the environment. Information systems are used to represent this partial information [23]. We distinguish two kinds of interactive attributes: sensory attributes and action attributes (Section 4). Our approach generalizes the concept of information systems known from the literature, from static information systems [13,14,15] to dynamic ones. It allows us for using information systems in modeling of interactive computations. There is growing research interest in the rough set community on dynamic information systems (see, e.g., [6,9]). Here, we emphasize that embedding this research in the framework of interactive computations is crucial for many real-life applications [8].

In Conclusions, we summarize the presented approach. We also give short information about our current research projects.

2 Interactive Computations

In this section, the global states are defined as pairs $(s_{ag}(t), s_e(t))$, where $s_{ag}(t)$ and $s_e(t)$ are states of a given agent ag and the environment e at time t, respectively. We now explain how the transition relation $\longrightarrow$ between global states are defined in the case of interactive computations. In Figure 1, the idea of transition from the global state $(s_{ag}(t), s_e(t))$ to the global state $(s_{ag}(t+\Delta), s_e(t+\Delta))$ is illustrated, where Δ is a time necessary for performing the transition, i.e., when $(s_{ag}(t), s_e(t)) \longrightarrow (s_{ag}(t+\Delta), s_e(t+\Delta))$ holds. $A(t)$, $E(t)$ denote the set of attributes available by agent ag at the moment of time t and the set of attributes used by environment e at time t, respectively. $Inf_{A(t)}(s_{ag}(t), s_e(t))$ is the signature of $(s_{ag}(t), s_e(t))$ relative to the set of attributes $A(t)$ and $Inf_{E(t)}(s_{ag}(t), s_e(t))$ is the signature of $(s_{ag}(t), s_e(t))$ relative to the set of attributes $E(t)$[1]. These signatures are used as arguments of strategies Sel_Int_{ag}, Sel_Int_e selecting interactions I_{ag} and I_e of agent ag with the environment and the environment e with the agent ag, respectively. $I_{ag} \otimes I_e$ denotes the result of the interaction product $\otimes$ on I_{ag} and I_e. Note that the agent ag can have very incomplete information about I_e as well as the result $I_{ag} \otimes I_e(s_{ag}(t+\delta), s_e(t+\delta))$ only, where δ denotes the delay necessary for computing the signatures and selection of interactions (for simplicity of reasoning we assume that these delays for a and e are the same). Hence, information perceived by a about $s_{ag}(t+\Delta)$ and $s_e(t+\Delta)$ can be very incomplete too. Usually, the agent ag can predict only estimations of $s_{ag}(t+\Delta)$ and $s_e(t+\Delta)$ during planning selection of the interaction I_{ag}. These predictions can next be compared with the perception of the global state $(s_{ag}(t+\Delta), s_e(t+\Delta))$ by means of attributes

[1] By considering only signatures over some set of attributes $E(t)$ we reflect one of the basic assumptions of interactive computing that interaction takes place in the environment which can not be controlled. $E(t)$ may not be known for an agent ag.

$A(t+\Delta)$. Note that $I_{ag} \otimes I_e$ can change the content of the agent state as well as the environment state. Assuming that the current set of attributes $A(t)$ is a part of the agent state $s_{ag}(t)$ this set can be changed, for example by adding new attributes discovered using I_{ag}, for example with the help of hierarchical modeling discussed previously. Analogously, assuming that the description of the strategy Sel_Int_{ag} is stored in the current state of the agent $s_{ag}(t)$ this strategy can be modified as the result of interaction. In this way, sets of attributes as well as strategies for selecting interactions can be adopted in time.

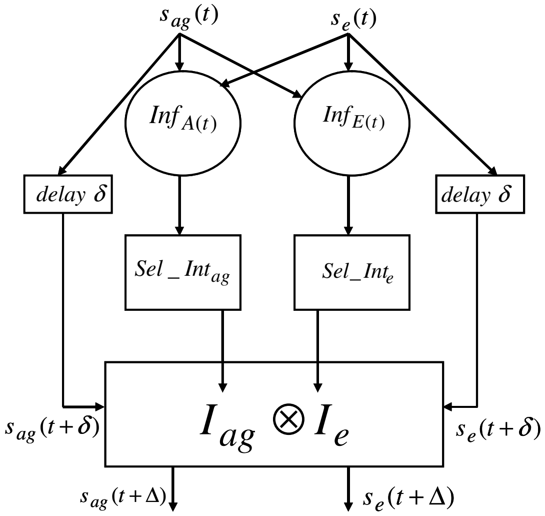


Fig. 1. Transition from global state $(s_{ag}(t), s_e(t))$ to global state $(s_{ag}(t+\Delta), s_e(t+\Delta))$

Computations observed by the agent ag using the strategy Sel_Int_{ag} in interaction with the environment e can now be defined with a help of the transition relation $\longrightarrow$ defined on global states and signatures of global states relative to the set of attributes of agent ag. More formally, any sequence $sig_1, \ldots, sig_n, ...$ is *a computation observed by ag in interaction with e* if and only if for some t, Δ and for any i, sig_i is the signature of a global state $(s_{ag}(t+i\Delta), s_e(t+i\Delta))$ relative to the attribute set $A(t+i\Delta))$ available by ag at a moment of time $t+i\Delta$ and $(s_{ag}(t+i\Delta), s_e(t+i\Delta)) \longrightarrow (s_{ag}(t+(i+1)\Delta), s_e(t+(i+1)\Delta))$[2].

Let us assume that there is given a quality criterion over a quality measure defined on computations observed by the agent ag and let sig_1 be a given signature (relative to the agent attributes). One of the basic problems for the agent ag is to discover a strategy for selecting interactions (i.e., selection strategy) in such a way that any computation (e.g., with a given length l) observed by ag and starting from any global state with the signature sig_1 and realized using the discovered selection strategy will satisfy the quality criterion to a satisfactory

[2] As usual one can consider finite and infinite computations.

degree (e.g., the target goal of computation has been reached or that the quality of performance of the agent ag in computation is satisfactory with respect to the quality criterion). The hardness of the selection strategy discovery problem by the agent ag is due to the uncertainty about the finally realized interaction, i.e., the interaction being the result of the interaction product on interactions selected by agent ag and the environment e. In planning the strategy, the agent ag can use (a partial) information on history of computation stored in the state. One may treat the problem as searching for the winning strategy in a game between the agent ag and the environment e with a highly unpredictable behavior.

3 Information Systems in Hierarchical Modeling

A hierarchical modeling of complex patterns (granules) in hierarchical learning (see e.g., [4]) can be described using the rough set approach based on information systems. In such description a construction of every model is described/made on the basis of a particular information system. The result of construction of an information system from a given level of hierarchical modeling is built from information systems from lower levels of its hierarchy. This is made by constructing of new attributes on the basis of already known attributes.

We use the standard rough set notation [13,14,15]. In particular, by $\mathcal{A} = (U, A, \{V_a\}_{a \in A})$, we denote an information system with the set of objects U, the set of attributes A, and the set of attribute values V_a for $a \in A$. By (U, C, d) we denote a decision system with the decision attribute d.

Attributes in information systems can be divided into two classes: *atomic attributes* and *constructible attributes*. Atomic attributes are basic in the sense that their values depend only on some external factors (with respect to a given information system) and are independent on values of other attributes. Atomic attributes can be both closed and open attributes.

Constructible attributes are complex attributes defined from other attributes, or more exactly inductively defined from atomic attributes. If b is a constructible attribute, then for any object x and already defined attributes $a_1, a_2, \cdots, a_m$: $b(x) = F(a_1(x), a_2(x), \cdots, a_m(x))$, where $F : V_{a_1} \times V_{a_2} \times ... \times V_{a_m} \longrightarrow V_b$ and elements of V_b are constructed on the basis of values values from V_i for $i = 1, .., m$.

Sensory attributes can serve as typical examples of atomic attributes. Values of sensory attributes are results of measurement conducted by sensors, so they depend only on the environment and are independent on values of other attributes. Constructible attributes defined from sensory attributes can represent higher-order results of perception, when some patterns are identified, perceptual granules are created *etc.*

We generalize the concept of attributes used in rough set theory [13,14,15,20]. In hierarchical modeling, the attribute value sets can be compound, i.e., they can be of higher order types represented, e.g., in the power set hierarchy [19,20]. The types are fixed in the case of atomic attributes and it should be properly constructed for constructible attributes. Note that elements of the attribute value sets can be complex objects such as signals or images. We also assume that for any attribute a together with its attribute value set V_a there is assigned a relational structure $\mathcal{R}_a = (V_a, \{r_i\}_{i \in I})$, where r_i are relations over Cartesian products of V_a. Examples of relational structures for attributes will be given later. Together with a relational structure $\mathcal{R}_a$ we consider a set of formulas $\mathcal{L}_a$ with interpretation in V_a, i.e., to any formula $\alpha \in \mathcal{L}_a$ there is assigned its meaning $\|\alpha\|_{\mathcal{R}_a} \subseteq V_a$. Moreover, for any attribute a there is distinguished a subset of formulas $L_a \subseteq \mathcal{L}_a$ defining a partition of V_a (defined using the semantics $\| \cdot \|_{\mathcal{R}_a}$). The result of interaction of any atomic attribute with the environment can be described as a selection of formula from L_a. Then values of attribute a, interpreted as the results of measurements by this attribute, can be identified with the index of the selected formula. One can observe that using this approach information systems can be interpreted as the result of a finite number of interactions with the environment of the relational structure defined by the Cartesian product of relational structures corresponding to attributes from information system.

Constructible attributes can be constructed in many ways [23]. One of them is based on introducing a relational structure on value domains of atomic attributes [23]. Because of the space limit we discuss one illustrative example only related to hierarchical modeling. Note that this modeling process is often hierarchical and interactions occur between constructed granules by a given agent on different hierarchical levels and between the levels. Granules are constructed on different levels by means of actions performed on granules from lower levels and then the quality of the granules constructed on higher levels is evaluated. In the case the quality is not satisfactory, new actions on higher levels should be activated in searching for construction of relevant granules. Observe that the concepts initiating actions are often drifting in time. Hence, the searching process (or the discovery process) for relevant granules should be adaptive. Moreover, the satisfiability of the approximated concepts is usually not binary but can be expressed as satisfiability to a degree only. Hence, mechanisms for conflict resolution between different concepts voting for initiating different actions should be developed/learned/discovered (analogously to voting mechanism between rules in the case of rule based classifiers). Another problem is related to propagation of satisfiability degrees of concepts through hierarchical levels. Here, the rough set approach proposes to use approximate reasoning schemes discovered from data or acquired from domain experts [4,18].

General operations on information systems can be defined as products with constraints (see [19] for more details). Definitely, one can consider at the next level of modeling sequences of time windows as structures and construct information systems or decision tables over such structural objects. Observe that in this case the indiscernibility (similarity) classes are sets of paths over time windows. One may, e.g., induce concurrent systems defining such sets of paths [21].

Discovery of relevant attributes on each level of the hierarchy is supported by domain knowledge provided e.g., by concept ontology together with illustration of concepts by means of samples of objects taken from this concepts and their complements [4]. Such application of domain knowledge often taken from human experts serves as another example of interaction of a system (classifier) with its environment. Additionally, for the support of relevant attributes, discovery on a given level as well as on other levels of the hierarchy can be found using different ontologies. These ontologies can be described by different sets of formulas and possibly by different logics. Note that in a hierarchical modeling of relevant complex patterns also top-down interactions of higher levels of hierarchy with lower levels should be considered, e.g., if the patterns constructed on higher levels are not relevant for the target task the top-down interaction should inform lower levels about necessity of searching for new patterns.

4 Interactive Attributes

There are two basic types of interaction between an agent and an environment: the influence of the environment on an agent and an opposite influence of an agent on its environment.

We need a specific class of attributes to represent interactions of an agent, namely *interactive attributes*, divided into two classes: *perception attributes* and *action attributes.*

Perception is one of the main forms of interaction of an agent with the environment. Moreover, this form is indispensable in the case of interactive systems. Without perception every action made by agent in the environment would be blind, without it agent would not be able to adapt its behavior to changing conditions of the environment or to modify dynamically its course of actions as a response to results of agent's actions in the environment.

In order to represent results of perception, we need a specific class of attributes: *perception attributes.* The beginning of the perception process is in senses in the case of living organisms or in sensors in the case of artificial agents. Senses/sensors interact with the environment. To represent the results of this interaction we use *sensory attributes.* These atomic attributes depend solely on interaction with the environment and are independent from other attributes in information system. Sensory attributes are also open attributes, i.e. if a is a sensory attribute, then a is a function with values in its value domain V_a. This reflects the fact that sensors interact with the environment which can not be controlled. Always it is possible that new stimuli appear to the senses/sensors which were not perceived before. The value domains of sensory attributes are determined only by sensitivity of sensors represented by these attributes.

In order to describe formally perception processes as interactions, let us introduce some notation. If $f : X \times Y \longrightarrow X \times Y$, then by $\pi_1[f]$, $\pi_2[f]$ we denote projections of f, i.e., $\pi_1[f] : X \times Y \longrightarrow X$, $\pi_2[f] : X \times Y \longrightarrow Y$ such that $f(x, y) = (\pi_1[f](x, y), \pi_2[f](x, y))$ for $(x, y) \in X \times Y$. A global state at time t, $s(t) = (s_{ag}(t), s_e(t))$, consists of the whole state of a given agent ag and its environment e. Let us recall that the agent ag does not have to posses complete information about these states and usually it does not. In Section 2, by I_{ag} and I_e we denote the influence operation of a given agent ag on its environment e and the opposite influence operation of the environment e on an agent ag, respectively. Both interactions can affect the global state of a given agent and its environment. By $I_e(s(t))$ $(I_{ag}(s(t)))$ we denote the global state at $t+\Delta$ obtained from $s(t)$ by applying I_e (I_{ag}) only. Since both I_{ag} and I_e last in time Δ they can also dynamically affect each other, a result of such interfering interaction is denoted by product $I_{ag} \otimes I_e$.

As we mentioned above, perception is an example of interaction between an agent and its environment. Moreover, it is a very interesting example. It is a kind of action made by an agent which usually do not affect the environment [3] but in which an agent is affected by its environment. In order to analyze perception process we should be more specific and introduce $I_{ag,a}$ - an interaction operation selected by an agent ag for performing measurement of the value of sensory attribute a. We assume that in $s_{ag}(t)$ are stored values of sensory attributes at time t, i.e., as a part of $s_{ag}(t)$ one can distinguish (a, v), where a is a sensory attribute and v is its value at time t (or information that this value is not available). In the described model changes of attribute values are recorded in discrete time timing with Δ. For a sensory attribute a we have that

$$\begin{aligned} s(t+\Delta) &= (s_{ag}(t+\Delta), s_e(t+\Delta)) = [I_{ag,a} \otimes I_e](s(t)) = \\ &= (\pi_1[I_{ag,a} \otimes I_e](s(t)), \pi_2[I_e](s(t))), \end{aligned} \tag{1}$$

assuming that: $s_{ag}(t+\Delta)$ differs from $s_{ag}(t)$ only on a part corresponding to attribute a, i.e., a new value of a is equal to the result of sensory measurement by $I_{ag,a}$ (in a more general case $s_{ag}(t)$ may be influenced by I_e) at time $t+\Delta$. Since $(I_{ag,a} \otimes I_e)(s(t)) = (\pi_1(I_{ag,a} \otimes I_e)[s(t)], \pi_2(I_{ag,a} \otimes I_e)[s(t)])$ therefore $\pi_2(I_{ag,a} \otimes I_e)[s(t)] = \pi_2[I_e(s(t))]$, i.e., $s_e(t+\Delta)$ was changed by I_e but there is no influence of $I_{ag,a}$. In other words $\pi_2[I_{ag,a}](s(t)) = I_e(s_e(t))$, i.e., $s_e(t+\Delta)$, the state of the environment e in time $t+\Delta$ being result of interaction is obtained from $s_e(t)$ by the dynamics of the environment only. In Figure 2, we illustrate the basic features of sensory attributes.

In the next steps (of perception), some new attributes can be introduced on the basis of information presented by sensory attributes in the ways described in Section 3. These are perception constructible attributes and we will refer to them as *complex perception attributes.* They correspond to complex perceptual representations constructed in the process of perception postulated in cognitive science [27], [2]. Complex perception attributes can be used in searching for patterns or structural properties of perceived objects (see Section 3). They also seem to be indispensable in solving classification problem for complex concepts in the case of newly perceived objects. Complex perception attributes serve as a kind of bridge between knowledge stored in an agent and results of perception given by sensory attributes. For the same reason, they are needed in approximation of complex vague concepts referring to environment perceived by a given agent and responsible for activating actions. Therefore complex perception attributes are also indispensable from the point of view of Wistech.

In rough set analysis of interactions, actions are represented by decision attributes. We refer to these attributes as *action attributes.* It follows from the

[3] In the case of quantum level it is not true.

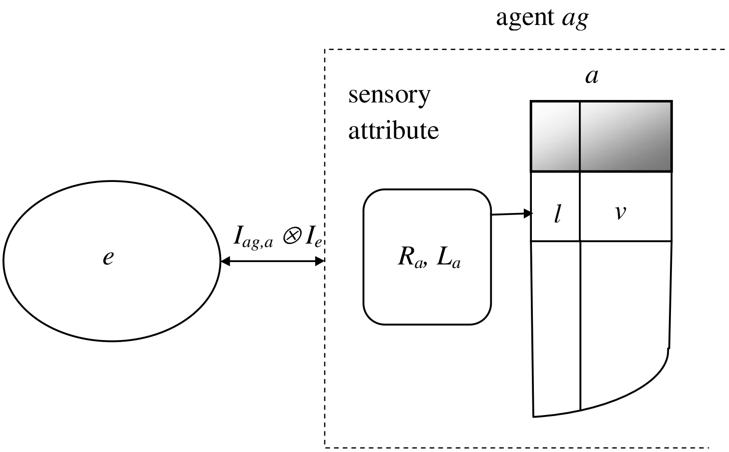


Fig. 2. Sensory attribute. e denotes the environment, $\mathcal{R}_a, L_a$ - relational structure of sensory attribute a and set of formulas assigned to a, respectively, l is a label of the environment state currently perceived by a, v is the index such that $\alpha_v \in L_a$ was selected in interaction of a with the environment. In the shadowed area the results of past interaction are stored, the interaction of a with the environment e is not changing e (the changes of e are caused by dynamics of the environment only). In the agent only a row with label l and v was added and represents the sensory attribute a measurement.

discussion made above, that action attributes should be somehow compound. A value of an action attribute should not only contain some information about elementary actions (or a chain of elementary actions - this difference is unessential) but also contain information about the specified goal and expected perceptual results of a given action/a chain of actions. These attributes can be constructed in many various ways. In the process Sel_Int_{ag}, i.e., the process leading to a selection of interaction $I_{ag,a}$, where an action attribute a is representing solely a given action/actions. The attribute a becomes also condition attribute and is used together with attributes representing knowledge of the agent ag and perception of ag for determining expected observable results in the environment. These anticipated results are compared with observable characteristics of specified goals and decisions about selection of interactions are made on the basis of their similarity or whether anticipated results match enough observable properties of goals. More advanced approach can use history of interaction for action prediction. Anticipated results of action predicted at time t can be compared to perceived states of the environment at time $t + \Delta$ being a result of interaction $[I_{ag,a} \otimes I_e](s(t))$. This comparison is used to modify an action in time $t + \Delta + \lambda$, where λ is time needed for making comparison and planning modification, in the case when perceived results are too far away from anticipated ones. The basic featutres of action attributes are illustrated in Figure 3.

In [23], an example of highly interactive cognitive architecture ACT-R [26] with a number of interactions between its different parts was discussed.

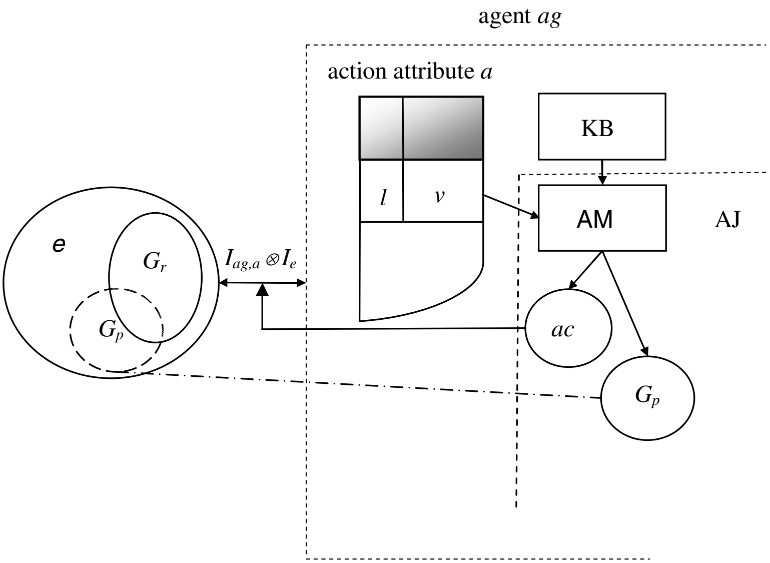


Fig. 3. Action attribute. On the basis of the current information about the current state of the agent ag and the state of the environment e, the action attribute a is selecting an action ac and predicts changes of the environment caused by ac which are represented by granule G_p. l, v have meaning as in Figure 2. AJ denotes the adaptive judgment module with the action submodule denoted by AM. The action attribute a is selecting an action ac to be performed (using module AM, knowledge base KB contents, and the measurement results stored by sensory attributes). Changes in e caused by ac in the form of granule G_p are predicted too. The selected action ac determines the interaction $I_{ag,a}$ with the environment from the side of the agent ag. Note that reaction of the environment may be unpredictable and the granule G_r representing change of e as the result of $I_{ag,a} \otimes I_e$ (on the component of the environment) may be different from predicted described by granule G_p.

5 Conclusions

We discussed the role of interactive computations in PBC. The fundamental role of information systems in modeling interactive computations was emphasized. Some more advanced representations of sensory measurements such as sets of information systems, clusters of information systems, or relational structures over information systems can be also considered. The role of these more complex structures for adaptive judgment and their relationships to the existing approaches (see, e.g., [3]) will be considered elsewhere.

We stressed the necessity of further development of the adaptive judgment methods for approximate reasoning about constructed granules as a crucial step in understanding of interactive computations. In particular, the necessity of using inductive strategies relevant for new features discovery, extension of approximation spaces, conflict resolution, tuning of quality measures, discovery of approximate reasoning schemes from data and domain knowledge, adaptive judgment based on beliefs, adaptive learning of concepts on the basis of histories of computations for prediction of actions or plans, hierarchy discovery (different levels with features and structural objects), reasoning by analogy, learning protocols for cooperation or competition, coalition formation are all example of tasks in which adaptive judgment is involved.

Acknowledgements. Authors would like to thank Professor Dominik Ślęzak for his critical comments helping to improve the paper. The research has been supported by the grants N N516 077837, N N516 368334 from Ministry of Science and Higher Education of the Republic of Poland.

References

1. Arbib, M.A.: The Metaphorical Brain 2: Neural Networks and Beyond. Willey & Sons, Chichester (1989)
2. Bara, B.G.: Cognitive Science. A Developmental Approach to the Simulation of the Mind. Lawrence Erlbaum Associates, Hove (1995)
3. Barwise, J., Seligman, J.: Information Flow: The Logic of Distributed Systems. Cambridge University Press, Cambridge (1997)
4. Bazan, J.: Hierarchical classifiers for complex spatio-temporal concepts. In: Peters, J.F., Skowron, A., Rybiński, H. (eds.) Transactions on Rough Sets IX. LNCS, vol. 5390, pp. 474–750. Springer, Heidelberg (2008)
5. Bower, J.M., Bolouri, H. (eds.): Computational Modeling of Genetic and Biochemical Networks. MIT Press, Cambridge (2001)
6. Chakraborty, M.K., Pagliani, P.: Geometry Of Approximation: Rough Set Theory: Logic, Algebra and Topology of Conceptual Patterns. Springer, Heidelberg (2008)
7. Goldin, D., Smolka, S., Wegner, P. (eds.): Interactive Computation: The New Paradigm. Springer, Heidelberg (2006)
8. Jankowski, J., Skowron, A.: Wisdom technology: A Rough-granular approach. In: Marciniak, M., Mykowiecka, A. (eds.) Bolc Festschrift. LNCS, vol. 5070, pp. 3–41. Springer, Heidelberg (2009)
9. Khan, M.A., Banerjee, M.: A study of multiple-source approximation systems. In: Peters, J.F., Skowron, A., Słowiński, R., Lingras, P., Miao, D., Tsumoto, S. (eds.) Rough Sets XII. LNCS, vol. 6190, pp. 46–75. Springer, Heidelberg (2010)
10. Maar, D.: Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. W.H. Freeman, New York (1982)
11. Mendel, J.M., Wu, D.: Perceptual Computing: Aiding People in Making Subjective Judgments. John Wiley & IEEE Press (2010)
12. Newell, A.: Unified Theories of Cognition. Harvard University Press, Cambridge (1990)
13. Pawlak, Z.: Rough sets. International Journal of Computing and Information Sciences 18, 341–356 (1982)
14. Pawlak, Z.: Rough sets. In: Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers, Dordrecht (1991)
15. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Science 177, 3–27 (2007); Rough sets: Some extensions. Information Science 177, 28–40 (2007); Rough sets and boolean reasoning. Information Science 177, 41–73 (2007)

16. Pedrycz, W., Skowron, A., Kreinovich, V. (eds.): Handbook of Granular Computing. John Wiley & Sons, Chichester (2008)
17. Poggio, T., Smale, S.: The mathematics of learning: Dealing with data. Notices of the AMS 50(5), 537–544 (2003)
18. Skowron, A., Stepaniuk, J.: Informational granules and rough-neural computing. In: Pal, S.K., Polkowski, L., Skowron, A. (eds.) Rough-Neural Computing: Techniques for Computing with Words, pp. 43–84. Springer, Heidelberg (2003)
19. Skowron, A., Stepaniuk, J.: Approximation spaces in rough-granular computing. Fundamenta Informaticae 100, 141–157 (2010)
20. Skowron, A., Stepaniuk, J.: Rough granular computing based on approximation spaces (Extended version of [19] submitted to the special issue of Theoretical Computer Science on Rough-Fuzzy Computing)
21. Skowron, A., Suraj, Z.: Discovery of concurrent data models from experimental tables: A rough set approach. In: Proceedings of First International Conference on Knowledge Discovery and Data Mining, pp. 288–293. AAAI Press, Menlo Park (1995)
22. Skowron, A., Wasilewski, P.: Information systems in modeling interactive computations on granules. In: Szczuka, M. (ed.) RSCTC 2010. LNCS, vol. 6086, pp. 730–739. Springer, Heidelberg (2010)
23. Skowron, A., Wasilewski, P.: Information systems in modeling interactive computations on granules (Extended version of [22] submitted to the special issue of Theoretical Computer Science on Rough-Fuzzy Computing)
24. Ślęzak, D., Toppin, G.: Injecting domain knowledge into a granular database engine – A position paper. In: CIKM 2010, Toronto, Ontario, Canada, October 26-30 (2010)
25. Sun, R.: Prolegomena to Integrating cognitive modeling and social simulation. In: Sun, R. (ed.) From Cognitive Modeling to Social Simulation, pp. 3–26. Cambridge University Press, Cambridge (2006)
26. Taatgen, N., Lebiere, C., Anderson, J.: Modeling paradigms in ACT-R 29. In: Sun, R. (ed.) Cognition and Multi-Agent Interaction. From Cognitive Modeling to Social Simulation, pp. 29–52. Cambridge University Press, Cambridge (2006)
27. Thagard, P.: Mind: Introduction to Cognitive Science, 2nd edn. MIT Press, Cambridge (2005)
28. Zadeh, L.A.: Computing with words and perceptions – A paradigm shift. In: Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI 2009), Las Vegas, Nevada, USA, IEEE Systems, Man, and Cybernetics Society (2009)
29. Zadeh, L.A.: Generalized theory of uncertainty (GTU) – principal concepts and ideas. Computational Statistics & Data Analysis 51(1), 15–46 (2006)
30. Zadeh, L.A.: Precisiated natural language (PNL). AI Magazine 25(3), 74–91 (2004)
31. Zadeh, L.A.: A new direction in AI: Toward a computational theory of perceptions. AI Magazine 22(1), 73–84 (2001)
32. Zadeh, L.A.: From computing with numbers to computing with words – From manipulation of measurements to manipulation of perceptions. IEEE Transactions on Circuits and Systems 45(1), 105–119 (1999)

Social Networks: Research and Applications

Ruay-Shiung Chang

Department of Computer Science and Information Engineering
National Dong Hwa University
Hualien, Taiwan

Abstract. On September 5, 2010, an opinion article in the Washington Post titled "Facebook and social media offer the potential of peace" appeared. The peace refers to the situation in Mideast. When young generations socialize between themselves using Facebook, MySpace, and/or Tweeter, war is less likely and peace may be possible.

Social media or social networks are changing the world, from how people make friends to how information is disseminated. More than 500 million people use Facebook alone. Of those, 70 percent are outside the United States. MySpace has 122 million monthly active users, and Twitter reports 145 million registered users. With their many users and vast amount of data produced each moment, social networks offer many research challenges in computer science.

For example, different from conventional data types, social media data are multi-model in nature, including content such as images, audio, videos, discussion topics, tags, annotations, hyperlinks, profiles, timestamps, click-throughs, etc. How to make meanings out of these data determines if intelligent applications can be devised.

The omnipresence of Internet makes social networks popular and easily accessible. Will social networks in turn affect the future network architecture? With the advancement of handheld mobile devices, social networks on the move are becoming common. Will mobile and pervasive social networks be any different?

Then there are trust, security, and privacy issues. As you can see, the list of related problems can go on and on and on. In this talk, we will identify some key issues in social network research and present some applications of social networks.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, p. 26, 2010.

The Roadmap for Sharing Electronic Health Records: The Emerging Ubiquity and Cloud Computing Trends

Sabah Mohammed and Jinan Fiaidhi

Department of Computer Science, Lakehead University
955 Oliver Road, Thunder Bay, Ontario P7B 5E1, Canada
{mohammed,jfiaidhi}@lakeheadu.ca

Abstract. Medical paper-based records have been in existence for decades and their gradual replacement by computer-based records has been slowly underway for over twenty years in healthcare systems. Computerized information systems have not achieved the same degree of penetration in healthcare as that seen in other sectors such as finance, transport and the manufacturing and retail industries. Further, deployment has varied greatly from country to country and from specialty to specialty and in many cases has revolved around local systems designed for local use. Electronic medical record systems lie at the center of any computerized health information system. Without them other modern technologies such as disease surveillance systems cannot be effectively integrated into routine clinical workflow. The paperless, interoperable, multi-provider, multi-specialty, multi-discipline computerized medical record, which has been a goal for many researchers, healthcare professionals, administrators and politicians for the past two decades, is however about to become reality in many countries. This article provides a roadmap vision based on the emerging web technologies that hold great promise for addressing the challenge of sharing electronic health records. It starts with addressing the ubiquity trend and how it can be realized based on the new cloud computing paradigm to share electronic heath records like the community of care records/documents (CCRs, CCDs). The article also addresses the security concerns related to sharing electronic health records over the cloud.

1 The Ubiquity Trend

While many countries have prioritized electronic health records (EHRs) as fundamental to increasing the efficiency and reducing the cost of healthcare, digital patient information is also foundational to improving the coordination and management of patient care, of transitioning from episodic care to managing wellness and ultimately to the delivery of personalized medicine. To make this transition, physicians, hospitals, insurance providers and pharmaceutical companies must be able to access, exchange and analyze patient information and that process must start with the digitization of patient information at the point of primary care. In spite of the technical improvements, the current healthcare systems often lack adequate integration among the key actors, and commonly fail to consider variety of interoperability and social aspects. Actually, there is a misconception that e-health is just about the usage of the

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 27–38, 2010.

Information and Communications Technologies (ICT). Actually, dealing with e-health applications, requires measures and technologies beyond the mere qualified communication networking infrastructures. The challenge will be to make the e-health technology as invisible as possible to attract widespread use. This means, it will need to be ubiquitous (i.e. present in every place) and widely accepted. For that to be true, there will need to be social as well as technological changes. Thus ubiquitous healthcare technologies become an emerging and challenging paradigm that is gradually reshaping the old disease-centered model, where treatment decisions are made almost exclusively by physicians based on clinical experience, into a patient-centered model where patients are active participants in the decision making process about their own health. Although the Internet has played a drastic role in this movement by giving people access to an extreme amount of health information and providing access to variety of e-health services, it fall short to present an effective media of participation and collaboration. With the other emerging unprecedented technological innovations (e.g. Wireless communication, cloud computing), many aspects of the e-health and healthcare systems are in need of serious modernization and fundamental shift. These current e-health systems are ripe with inefficiencies, inequities and errors. Additionally, the growing adoption of electronic health records systems and the steady accumulation of patient data at hospitals and clinics place an ever-increasing demand on digital storage and associated security issues. Thus, a new perspective on e-health is more critical than ever before. There are three major elements that can contribute to the ubiquity of e-health services (see Figure 1).

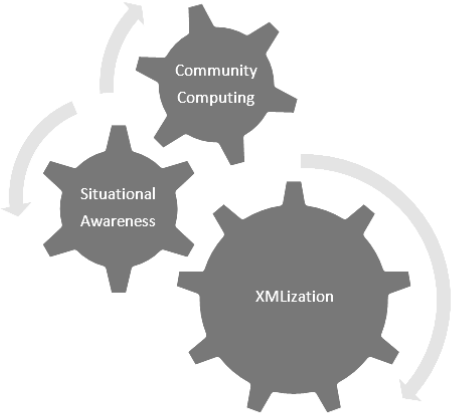


Fig. 1. Elements of the Ubiquity Trend

One of the first problems facing e-health ubiquity and interoperability is the fact that data are invariably stored in ASCII plain text format. This format does not reflect the underlying structure of the data. Consequently, e-health software tools need to unearth this structure for ubiquity and interoperability purposes. For this purpose, efforts around the world have been taken to explore the possibility of adopting XML format to represent the structure of e-health data. Such efforts try to employ standards that will enable consumers and e-health systems to more accurately search information on the Web as well as to interpret one form of XML-based document to another.

Actually XMLization can be devided into two layers: *Syntactic XMLization* and the *Semantic XMLization*. Syntactic XMLization defines the messaging layer and involves the ability of two or more systems to exchange information. Syntactic XMLization involves several sub-layers: network and transport layer (such as Internet), application protocol layer (such as HTTP or email), messaging protocol and message format layer (such as ebXML messaging or SOAP), and the sequencing of the messages. Syntactic XMLization guarantees the message to be delivered but does not guarantee that the content of the message will be machine processable at the receiving end. To guarantee message content interoperability, the content should conform to an agreed semantic. Semantic XMLization is the ability for information shared by systems to be understood at the level of formally defined domain concepts. Around the world, considerable gains have been achieved already, through standardization based on syntactic XMLization. The requirements of semantic XMLization far exceed those of syntactic XMLization, requiring a stable reference information model, vocabulary bindings to controlled terminologies, formally defined data types and structures, a mechanism for defining and constraining clinical statements and documents, a common repository of consensus-based reusable clinical concepts, and an agreed interchange format. Currently the only viable candidates for this role are HL7 version 3 (with or without CDA), openEHR and CEN 13606. Yet, it is not realistic to expect all the healthcare institutes to conform to a single standard. Therefore, there is a need to address the ubiquity and interoperability problem at the semantic level.

At the forefront of the semantic XMLization is the issue of semantic translation; that is, the ability to properly interpret the elements, attributes, and values contained in an XML file. In many cases, specific healthcare domains have standardized the way data are represented in XML. When this does not occur, some type of mediation is required to interpret XML formatted data that does not adhere to pre-defined semantics. Although this mediation and interpretation process is at the heart of the ``schema-wars" which are currently raging at forums such as www.w3c.org, www.oasis-open.org, www.rosettanet.org, www.schema.net, etc., this fundamental aspect of XML is not yet widely recognized or solved. Such mediation is actually needed before data exchange can be established between applications where a model of the domain of interest has to be built, since it is necessary to clarify what kind data a sent from the first application to the second. This model is usually described in terms of objects and relations. From the domain model a DTD or an XML Schema is constructed. Since the DTD or the XML Schema is just a grammar, then there exist multiple possibilities to encode a given element. To achieve semantic XMLization, one need to have constrained the specifications of XML Schemas/DTD so that a single translation mechanism can be used to transform any XML instances compliant with Ontology-based Naming and Design Rules into standard RDF or OWL structures. In this way, ontologies can be used to integrate, interoperate and mediate information exchange. However, building such mediators is extremely complicated process even for different versions of the same healthcare system. In this direction, decision-making is a central cognitive activity required for this type of mediators. Many aspects of EHR-based decision-making (e.g. diagnosis, and choice of therapy and tests) have been described using naturalistic decision-making [1], an emerging area of cognitive study which recognizes that expert decisions are frequently based on pattern recognition and situation awareness rather than formal logic. Both modes of decision making can be supported by the EHR. Rule-based decision making requires "facts" from the EHRs: labs, demographics, studies and events. Naturalistic decision making relies on decisions cued by patterns of facts, many of which are expressed in the narrative text that is written by the clinicians, doctors or nurses, during patient care and found in notes, reports and summaries in the EHR. In both cases, decision making based on an electronic patient chart becomes increasingly complex as the volume of data in the e-health system increases [2]. Coping with the growth of data in successfully implemented EHR systems is an emerging cognitive challenge for EHR users. In this direction, statistical methods can be used for identify similarities and differences based on sound statistical software like the SatScan.[1] However, the capacity and intelligence of newly developed e-health systems are growing day by day. For highly complex problems requiring diverse capabilities, an approach based on cooperation among elements can be an efficient solution [3]. Because such a cooperative based approach involves continuous and rich interactions, multi-agent technology is frequently used to design and develop cooperation based ubiquitous e-health systems. In addition, agents' characteristics such as intelligence and autonomy are suitable for developing intelligent ubiquitous computing systems that can adapt to dynamically changing situations. Jung et al. [4] propose Community Computing as an agent-based development paradigm for ubiquitous computing systems. The objective of community computing framework is to provide ubiquitous services through dynamic cooperation among agents. This approach focuses more on cooperation compared to the other multiagent methodologies.

2 The Cloud Computing Trend

Before embarking on a discussion of the possible directions for realizing the ubiquity trend, it is important to know the key e-health enhancement drivers [5]:

- **Reducing the Delivery of Cost of Care:** There is an international trend of increasing healthcare costs due to the ageing population, increasing burden of chronic disease, and increasingly expensive care options, and wastage of resources through inappropriate, unnecessary or duplicated interventions.
- **Comprehensive Availability and Ubiquity:** As most of the e-health services are available at urban areas.
- **Providing Context-Awareness:** There is a need for e-health services that are aware of the patient's presence and their context. Such awareness need to be sensitive, adaptive, and responsive to their needs, habits, gestures and emotions.
- **Empowerment of e-health Providers and Consumers:** Most of the e-health services are confined to the providers servers with very little empowerment on the consumer side
- **Providing Patient Safety:** Significant diagnostic and therapeutic error rates are observed in all jurisdictions based on the previous IT developments.

[1] `http://www.satscan.org/`

- **Synchronization of Content and Application Development:** Development efforts of e-health services are typically uncoordinated and essentially independent.
- **Integration of e-health Segments:** There is a need to integrate the various features and functions of e-health tools, including health information and support, transaction processing, electronic health records, clinical and public health information systems, compliance and disease management programs, and behavior change and health promotion.
- **Encourage New Types of Collaborations:** E-health services need to provide many new collaboration relationships (e.g. provider to provider, virtual teams).
- **Enhancing the Continuity and Coverage of Care:** Patients are increasingly seen by an array of providers in a wide variety of organizations and places, raising concerns about fragmentation of care. Healthcare needs to have wider coverage too (e.g. pervasive, mobile, RFID-based).
- **Support Sound Research and Education:** There is increasing level of demand from various healthcare sections (e.g. pharmaceutical industry, genomic research) to have e-health research data available and accessible based on standard format for research and education purposes.

By adhering to these drivers we can provide better e-health services in underserved areas, networked health systems and patients can gain access to seamless, coordinated and continuing care. This vision requires very much an innovative information technology infrastructure on the Internet that encourages collaboration, flexibility and integration. The cloud computing trend is the backbone that can take the abstract vision of the ubiquity trend to the concrete reality. The cloud computing [6] promises reliable services delivered through next-generation data centers that are built on compute and storage virtualization technologies. Consumers will be able to access applications and data from a "Cloud" anywhere in the world on demand. In other words, the Cloud appears to be a single point of access for all the computing needs of consumers. The consumers are assured that the Cloud infrastructure is very robust and will always be available at any time. The cloud computing enterprise relies on the Internet for satisfying the computing needs of the users. Cloud computing services usually provide common business applications online that are accessed from a web browser, while the software and data are stored on the servers. Cloud services - computation services, storage services, networking services, whatever is needed - are delivered and made available in a simplified way - "on demand" regardless of where the user is or the type of device they're using [7]. It enables both rapid innovation and support of core business functions based on technologies that are termed as SaaS (Software as Service), dSaaS (Data Storage as Service), PaaS (Platform as Service) and IaaS (Infrastructure as Service) [8]. Currently there are many venders who offer such services: (Google Apps, Microsoft "Software+Service"—SaaS), (Nirvanix SND, Amazon S3—dSaaS), (IBM IT Factory, Google AppEngine—PaaS) and (Amazon EC2, IBM BlueCloud—IaaS). Any cloud computing architecture can have a wide variety of reusable components, including portals, mashup servers, RSS feeds, database servers, mapping servers, gadgets, user interface components, and legacy Web servers. Architects can integrate these elements to form an edge application in which multiple services and components are overlaid or "mashed" together [9]. Actually, cloud computing offers a path to an enterprise IT infrastructure and application portfolio that can be both radically simple and more effective than the burdensome patchwork of clunky legacy systems that you find in enterprises today. In order to share EHRs over the cloud we need to employ the notions described in the following four subsections [10].

2.1 EHRs Frontend Portal

The frontend portals allow applications from a variety of different mashups, widgets and iframes to be included in different user panels. Such frontend portals also let users build applications through simple customization. The frontend portals represent a software package that includes browsers, media players, word processing programs, and spreadsheet software. They can display all kinds of information to healthcare users, including narrative text, graphics, images, video, audio, structured or tabular data, charts, highly structured documents, electronic forms, etc.

2.2 EHRs SaaS Backend

The backend can be in the form of SaaS services used by a local application or a remote application observed through a Web browser. Remotely executing applications commonly rely on an application server to expose needed services. An application server (e.g. Red Hat JBoss, Apatche Geronimo) is a software framework that exposes APIs for software services (such as transaction management or database access). Among the most important SaaS services are the followings:

- **UI SaaS:** In order to provide rich user interface controls to craft high-end fat client application.
- **Gadget SaaS:** Allow developers to easily write useful web applications that work anywhere on the web without modification.
- **Mashup SaaS:** To provide mashup and customization services that enable the development of mini applications.

2.3 EHRs Data as XML

When data is in XML, thousands of existing software packages can handle that data. Data in XML is universally approachable, accessible, and usable. Thus the new standards of sharing EHRs are based on XML. There are two notable standards for sharing EHRs over clouds: the Google Health Community of Care Record (CCR) Standard and Microsoft HealthVault Community of Care Document (CCD) standard. The CCR was created by ASTM with the objective to establish a standard that would be far easier to deploy and use by smaller physician practices. The CCD standard was created by the HL7 to serve large hospitals and healthcare institutions. Both standards are meant to ease the exchange of clinical information with a relatively easy to read and practical data-format and schema. You may convert one format to another (e.g. CCR to CCD mappers[2]). The XML format of CCR is very straight forward, consisting of a header, body and footer with the following top-level elements (Figure 2) [11]:

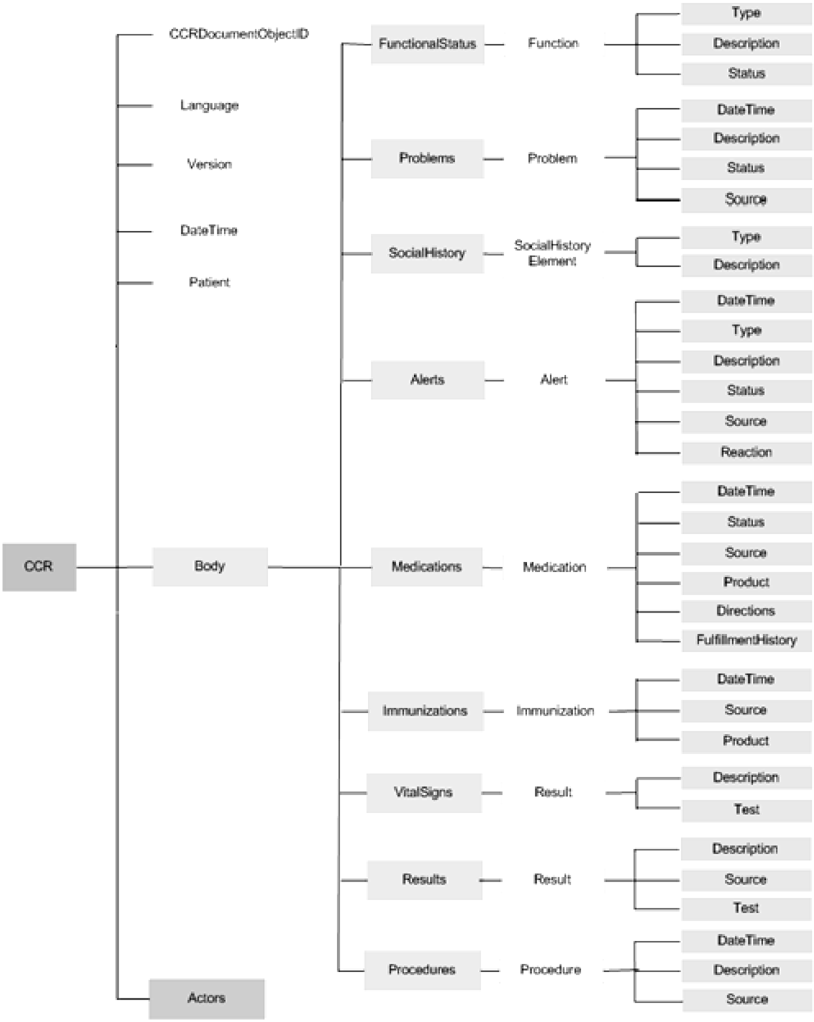
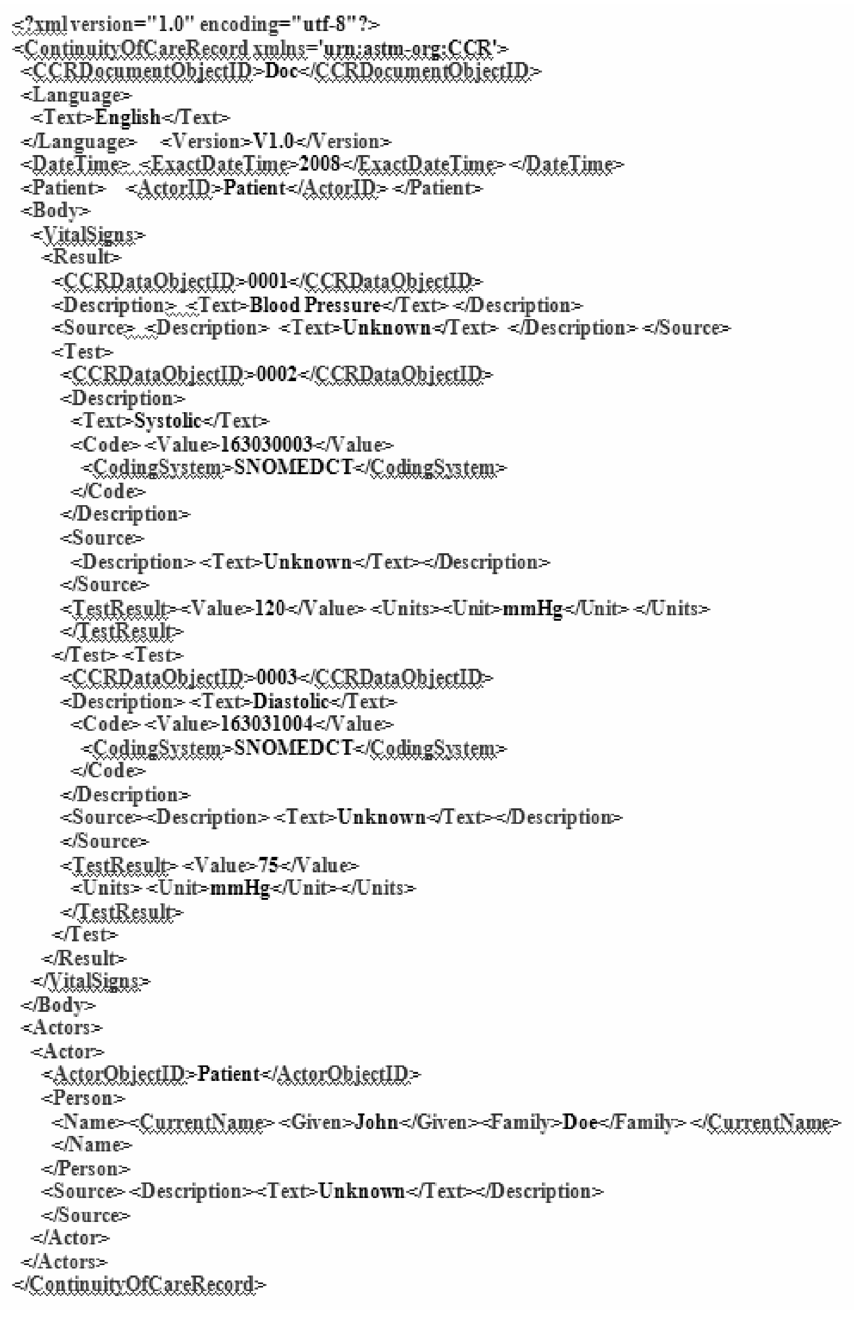


Fig. 2. The CCR Format

For example, using the SNOMED-CT terminologies one can write the Systolic Blood pressure reading in XML CCR format as in Figure 3 [12]:

[2] `http://wiki.hl7.org/index.php?title=Mappings_and_Translations`



```
<?xml version="1.0" encoding="utf-8"?>
<ContinuityOfCareRecord xmlns='urn:astm-org:CCR'>
 <CCRDocumentObjectID>Doc</CCRDocumentObjectID>
 <Language>
   <Text>English</Text>
 </Language>    <Version>V1.0</Version>
 <DateTime>  <ExactDateTime>2008</ExactDateTime> </DateTime>
 <Patient>    <ActorID>Patient</ActorID> </Patient>
 <Body>
   <VitalSigns>
     <Result>
       <CCRDataObjectID>0001</CCRDataObjectID>
       <Description>  <Text>Blood Pressure</Text> </Description>
       <Source>  <Description>  <Text>Unknown</Text>  </Description> </Source>
       <Test>
        <CCRDataObjectID>0002</CCRDataObjectID>
        <Description>
          <Text>Systolic</Text>
          <Code> <Value>163030003</Value>
            <CodingSystem>SNOMEDCT</CodingSystem>
          </Code>
        </Description>
        <Source>
          <Description> <Text>Unknown</Text></Description>
        </Source>
        <TestResult><Value>120</Value> <Units><Unit>mmHg</Unit> </Units>
        </TestResult>
       </Test> <Test>
        <CCRDataObjectID>0003</CCRDataObjectID>
        <Description> <Text>Diastolic</Text>
          <Code> <Value>163031004</Value>
            <CodingSystem>SNOMEDCT</CodingSystem>
          </Code>
        </Description>
        <Source><Description> <Text>Unknown</Text></Description>
        </Source>
        <TestResult> <Value>75</Value>
          <Units> <Unit>mmHg</Unit></Units>
        </TestResult>
       </Test>
     </Result>
   </VitalSigns>
 </Body>
 <Actors>
   <Actor>
     <ActorObjectID>Patient</ActorObjectID>
     <Person>
       <Name><CurrentName> <Given>John</Given><Family>Doe</Family> </CurrentName>
       </Name>
     </Person>
     <Source> <Description><Text>Unknown</Text></Description>
     </Source>
   </Actor>
 </Actors>
</ContinuityOfCareRecord>
```

Fig. 3. An Example of CCR

2.4 EHRs Processing and Awareness Protocols

The EHRs awareness protocols allow client applications basically to discover view and update CCR/CCD records in the form of data feeds. Based on these protocols, the client application can create new entries, edit or delete existing entries, request a list of entries, and query for entries that match particular criteria. These protocols can be simple where it uses the raw XML CCR/CCD data along with HTTP to process EHRs from the cloud repository. For this purpose developers can use available APIs like the Google Data API[3]. However, awareness protocols may have variety of additional functionalities such as:

- **EHRs Harvesting:** The ability to efficiently harvest patient information from EHRs repositories. This capacity gives rise to a new suite of potential uses for patient health information, as it can form the basis of meaningful quality assurance and accountability processes, as well as powerful epidemiological studies [13].*There are variety of APIs that can be used for this purpose (eg. Web-Harvest*[4], NEDLIB Harvester[5], HTTrack[6]).
- **EHRs Mashups:** With the rise in the use of the web for health purposes, patients have begun to manage their own health data online, use health-related services, search for information, and share it with others. The cooperation of healthcare constituents towards making collaboration platforms available is known today as Health 2.0 [14]. The significance of Health 2.0 lies in the transformation of the patient from a healthcare consumer to an active participant in a new environment. The basic protocol for enabling the integration of relevant data, services, and applications is unknown as the mashup (e.g. Medic-kIT [15]) which adheres to Web 2.0 standards.

3 The Security Concerns of Sharing EHRs over the Cloud

As always, working in a new environment presents new and different challenges for developers as well as users. Concern about security is one of the greatest hurdles to implementing the notions of ubiquity trend for the e-health cloud paradigm. Privacy and security problems are compounded by the use of remote, distributed services operated by third parties. E-health institutions employing these emerging technologies must take a comprehensive approach to security across on-site and cloud infrastructure, and learn how to assess and manage it. This encompasses protection, access and management, all built around user identity and integrated with a highly secure, interoperable platform for a broad set of partner solutions. Identity is a core part of any security strategy, because it allows for more contextual protection and access to information and resources. Identity is important to identify the user for different purposes including personalization, authorization and communication, content publishing and maintaining public identity across different e-health providers. In general, the identity management systems are elaborated to deal with the following core facets [16]:

[3] http://code.google.com/apis/gdata/

[4] http://web-harvest.sourceforge.net/

[5] http://www.csc.fi/sovellus/nedlib

[6] http://www.httrack.com

- **Management:** The amount of digital identities per person will increase, so the users need convenient support to manage these identities and the corresponding authentication.
- **Reachability:** The management of reachability allows users to handle their contacts to prevent misuse of their address (spam) or unsolicited phone calls.
- **Authenticity:** Ensuring authenticity with authentication, integrity and non-repudiation mechanisms can prevent identity theft.
- **Anonymity and Pseudonymity:** Providing anonymity prevents tracking or identifying the users of a service.
- **Organization of personal data management:** A quick method to create, modify or delete work accounts is needed, especially in big organizations.

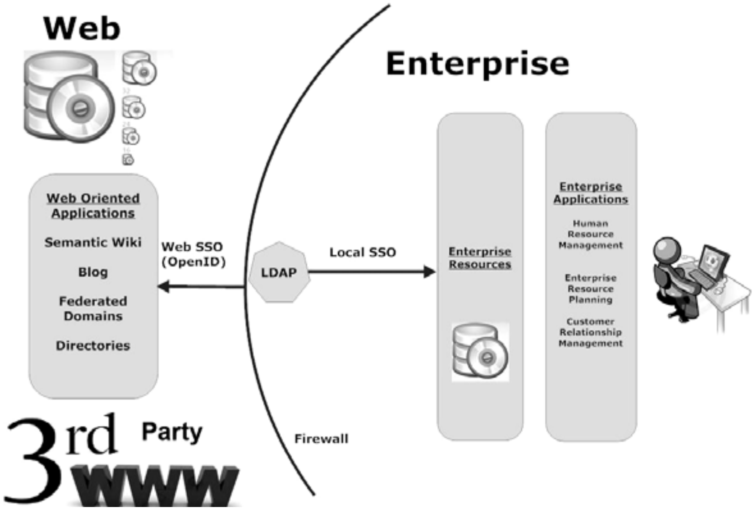


Fig. 4. Enterprise versus Web Identity Management

The roadmap of identity management started with enterprise identity management but later been developed into more general management that is currently known as user-centric Identity Management [17]. On one hand, enterprise identity management provides centralized security and authentication that is closed to third parties. On the other hand, the user centric identity management is based on decentralized authentication but it allows users to keep at least some control over their personal data. With the advancement in user-centric and URI-based identity systems over the past few years for accessing open applications and services on the web, it has become clear that a single specification will not be the solution to all security problems. Rather, like the other layers of the Internet, developing small, interoperable specifications that are independently implementable and useful will ultimately lead to market adoption of these technologies. Currently, there are wave of specifications, tools and techniques that are termed “identity 2.0” [18] to deal with sharing applications and services on the web and outside the boundary of the enterprise. Identity 2.0 is the anticipated revolution of identity verification on the Internet using emerging user-centric technologies such as OpenID[7], Open SSO[8], Open SAML[9], Live ID[10] and OAuth[11]. Identity 2.0 stems from the Web 2.0 initiative of the World Wide Web transition. Its emphasis is a simple and open method of identity transactions similar to those in the physical world, such as driver's license. The goal is to centralize security around users instead of organizations. Figure 4 illustrates the two main themes of identity management that can be used for cloud computing applications.

4 Conclusion

Recent foray by lead vendors like Amazon, Google, IBM, Sun Microsystems, Microsoft etc. into cloud computing, advocate for a universal sharing scenario where enterprises are not required to own any infrastructure or applications of their own. This is enabled by powerful usage of service-oriented architectures - namely infrastructures available as a set of usable pluggable services. Actually, cloud computing offers an ideal alternative to heavy investment in server and hardware infrastructure that most information systems require, as well as an easy means of scaling resources to meet changes on demand. In periods of economic uncertainty institutions look at the cloud computing option to enable them to continue to meet end-user and business demands for IT services. Cloud computing is an opportunity for business to implement low cost, low power and high efficiency systems to deliver scalable infrastructure. It increases capacity and expands computing capabilities without heavy investment in infrastructure, training or software licensing [19].

Triggered by an aging and populace, rising population numbers, and an increased awareness of the need to create a continuum of care amongst health community and healthcare institutions, users now desire quick and easy access to accurate patient data. Cost and poor usability have been cited as the biggest obstacles to adoption of health IT – especially the EHR systems – and have resulted in problematically-low EHR adoption rates. Cloud computing may help in eliminating this cost, and IT maintenance burdens that are often beyond the reach of small medical practices. Elimination of barriers to EHR adoption such as costly hardware infrastructure, while at the same time ensuring security and privacy of protected health information, are all positive results of using the “cloud” approach [20]. This article outlines the ubiquity trend and how cloud computing can implement the notion of ubiquitous systems for sharing EHRs. It addresses also the security concerns of sharing EHRs over the cloud.

References

1. Hoffman, R.R., Militello, L.G.: Perspectives on cognitive task analysis: Historical origins and modern communities of practice. Psychology Press, New York (2009)
2. Horsky, J., Zhang, J., Patel, V.L.: To err is not entirely human: Complex technology and user cognition. J. Biomed. Inform. 38(4), 264–266 (2005)

[7] http:// openid.net/

[8] https://opensso.dev.java.net

[9] http://OpenLiberty

[10] www.passport.net/

[11] http://oauth.net

3. Wooldridge, M., Jennings, N.R.: The Cooperative Problem-Solving Process. Journal of Logic Computation 9(4), 563–592 (1999)
4. Jung, Y., Lee, J., Kim, M.: Community Computing Model supporting Community Situation based Cooperation and Conflict Resolution. In: Obermaisser, R., Nah, Y., Puschner, P., Rammig, F.J. (eds.) SEUS 2007. LNCS, vol. 4761, pp. 47–56. Springer, Heidelberg (2007)
5. Eng, T.R.: The eHealth Landscape: A Terrain Map of Emerging Information and Communication Technologies in Health and Health Care. The Robert Wood Johnson Foundation, Princeton (2001)
6. Weiss, A.: Computing in the Clouds. Networker 11(4), 16–25 (2007)
7. Jones, M.T.: Cloud computing with Linux Cloud computing platforms and applications, IBM Research Journal (2008), http://www.ibm.com/developerworks/library/l-cloud-computing/index.html (retrieved on April 25, 2009)
8. Kaliski, B.: Multi-tenant Cloud Computing: From Cruise Liners to Container Ships. In: Third Asia-Pacific Trusted Infrastructure Technologies Conference (APTC), p. 4 (2008)
9. Hutchinson, C., Ward, J., Castilon, K.: Navigating the Next-Generation Application Architecture. IT Professional 11(2), 18–22 (2009)
10. Mohammed, S., Fiaidhi, J. (eds.): Ubiquitous Health and Medical Informatics: The Ubiquity 2.0 Trend and Beyond. IGI Global (2010) ISBN13: 9781615207770
11. Waldren, S.: Introduction to CCR Standard and XML Schema. Video presentation, http://www.veoh.com/collection/astmccr/watch/v14141513WQRzgjzc
12. Conner, K.: Understanding CCR. Health blog posting, http://healthblog.vitraag.com/2009/10/understanding-ccr/
13. Longstaff, D.: Contentious Crop: Harvesting Information from Electronic Health Records, Australian National University (2005), http://www.anu.edu.au/aphcri/Publications/Duncan_Longstaff_Internship_Paper-Harvesting_Information_From_Electronic_Health_Records.pdf
14. Van De Belt, T.H.: Definition of Health 2.0 and Medicine 2.0: A Systematic Review. J Med. Internet Res. 12(2), e18 (2010)
15. Greenshpan, O., et al.: Towards Health 2.0: Mashups to the rescue. In: 7th Int. Conf. on Next Generation Information Technologies and Systems, pp. 63–72 (2009)
16. El Maliki, T., Seigneur, J.M.: A Survey of User-centric Identity Management Technologies. In: Int. Conf. on Emerging Security Information, Systems and Technologies, IEEE SECUREWARE (2007)
17. Hansen, M.A.: User-controlled identity management: The key to the future of privacy? International Journal of Intellectual Property Management 2(4), 325–344 (2008)
18. Hamlin, K.: Identity 2.0 Gathering: Getting to the Promised Land (2005), http://www.oreillynet.com/pub/a/policy/2005/10/07/identity-workshop.html (retrieved on January 25, 2009)
19. Geelan, J.: The Cloud Computing Ecosystem: The Top 100 Cloud Players. Journal of Cloud Computing (January 13, 2009)
20. Rowley, R.: Is Cloud Computing Right for Health IT? EHR Bloggers, August 6 (2009), http://www.ehrbloggers.com/2009_08_01_archive.html

Compound Analytics of Compound Data within RDBMS Framework – Infobright's Perspective

Dominik Ślęzak[1,2]

[1] Institute of Mathematics, University of Warsaw Banacha 2, 02-097 Warsaw, Poland
[2] Infobright Inc., Krzywickiego 34 lok. 219, 02-078 Warsaw, Poland
slezak@infobright.com

The relational model has been present in research and applications for decades, inspiring a number of RDBMS products based on entirely different architectures, but sharing the same way of understanding and representing the data [4]. Given 40 years of history, it is clear that the relational paradigms should not be blindly followed in all situations [1]. On the other hand, given its popularity, the relational framework is usually the easiest one to accept by database users and the most convenient for interfacing with other tools.

An important trend in database industry relates to *analytical engines* that are optimized for advanced reporting and ad hoc querying. Such engines are usually applied at the level of data marts, especially in market segments where rapid data growth is expected. Originally, they have been technically complex and difficult to maintain. However, they have evolved toward solutions such as, e.g., Infobright's Community/Enterprise Editions (ICE/IEE)[1], capable of handling tens of terabytes of data on a single off-the-shelf box [15].

Infobright's engine is a fully functional RDBMS product with external connectors provided via integration with MySQL, and internals based on columnar storage [8], adaptive compression [17], as well as compact *rough* information that replaces standard database indexes [13]. We refer, e.g., to [3,10,16] for current research on ICE/IEE core technology, and to [2,7,12] for several interesting examples of its usage in academic and commercial projects.

In this talk, we use Infobright's software as a baseline to discuss limitations of relational model with respect to modern database applications. In particular, we investigate some challenges related to *compound analytics* and *compound data.* In both cases, we claim that it would be a mistake to give up too quickly the benefits of a typical RDBMS way of interacting with users. Instead, we present some application-level and technology-level solutions that do not contradict with original relational framework's universality and simplicity.

With regards to compound analytics, as an example, we consider practical inspirations and opportunities for enriching standard SQL language with approximate aspects [5,11], assuming minimum impact on query syntax and maximum easiness of interpreting inexact query answers.

With regards to compound data, we discuss two general approaches to employ domain knowledge about data semantics in order to improve database efficiency:

[1] www.infobright.{org.com}

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 39–40, 2010.

1) expressing data hierarchies explicitly at data schema level (see e.g. [6,12]), or 2) doing it independently from both logical and physical modeling layers, taking into account that domain experts may need interfaces other than those designed for database end-users and administrators (see e.g. [9,14]).

References

1. Agrawal, R., et al.: The Claremont report on database research. SIGMOD Rec. 37(3), 9–19 (2008)
2. Apanowicz, C.: Data Warehouse Discovery Framework: The Case Study. In: Zhang, Y., et al. (eds.) DTA/BSBT 2010. CCIS, vol. 118, pp. 159–170. Springer, Heidelberg (2010)
3. Borkowski, J.: Performance debugging of parallel compression on multicore machines. In: Wyrzykowski, R., et al. (eds.) PPAM 2009. LNCS, vol. 6068, pp. 82–91. Springer, Heidelberg (2010)
4. Codd, E.F.: Derivability, redundancy and consistency of relations stored in large data banks. SIGMOD Rec. 38(1), 17–36 (2009) (Originally: IBM Research Report RJ599, 1969)
5. Cuzzocrea, A.: OLAP Data Cube Compression Techniques: A Ten-Year-Long History. In: Kim, T.-h., et al. (eds.) FGIT 2010. LNCS, vol. 6485, pp. 751–754. Springer, Heidelberg (2010)
6. Das, S., Chong, E.I., Eadon, G., Srinivasan, J.: Supporting ontology-based semantic matching in RDBMS. In: Proc. of VLDB 2004, pp. 1054–1065. Morgan Kaufmann, San Francisco (2004)
7. Frutuoso Barroso, A.R., Baiden, G., Johnson, J.: Knowledge Representation and Expert Systems for Mineral Processing Using Infobright. In: Proc. of GRC 2010, pp. 49–54. IEEE, Los Alamitos (2010)
8. Hellerstein, J.M., Stonebraker, M., Hamilton, J.R.: Architecture of a Database System. Foundations and Trends in Databases 1(2), 141–259 (2007)
9. Moss, L.T., Atre, S.: Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications. Addison-Wesley, Reading (2003)
10. Ślęzak, D., Kowalski, M.: Intelligent Data Granulation on Load: Improving Infobright's Knowledge Grid. In: Lee, Y.-h., et al. (eds.) FGIT 2009. LNCS, vol. 5899, pp. 12–25. Springer, Heidelberg (2009)
11. Ślęzak, D., Kowalski, M.: Towards Approximate SQL – Infobright's Approach. In: Szczuka, M., et al. (eds.) RSCTC 2010. LNCS, vol. 6086, pp. 630–639. Springer, Heidelberg (2010)
12. Ślęzak, D., Sosnowski, Ł.: SQL-Based Compound Object Comparators – A Case Study of Images Stored in ICE. In: Kim, T.-h., et al. (eds.) ASEA 2010. CCIS, vol. 117, pp. 304–317. Springer, Heidelberg (2010)
13. Ślęzak, D., Synak, P., Wróblewski, J., Toppin, G.: Infobright – Analytic Database Engine using Rough Sets and Granular Computing. In: GRC 2010, pp. 432–437. IEEE, Los Alamitos (2010)
14. Ślęzak, D., Toppin, G.: Injecting Domain Knowledge into a Granular Database Engine – A Position Paper. In: Proc. of CIKM 2010, pp. 1913–1916. ACM, New York (2010)
15. Ślęzak, D., Wróblewski, J., Eastwood, V., Synak, P.: Brighthouse: An Analytic Data Warehouse for Ad-hoc Queries. PVLDB 1(2), 1337–1345 (2008)
16. Synak, P.: Rough Set Approach to Optimisation of Subquery Execution in Infobright Data Warehouse. In: Proc. of SCKT 2008. PRICAI 2008 Workshop (2008)
17. Wojnarski, M., et al.: Method and System for Data Compression in a Relational Database. US Patent Application 2008/0071818 A1 (2008)

From Grid Computing to Cloud Computing: Experiences on Virtualization Technology

Hai Jin

Huazhong University of Science and Technology
Wuhan, 430074, China
hjin@hust.edu.cn

Abstract. In recent years, we see a great change from grid computing to cloud computing. Cloud computing is among the most popular words in computing today. As technology foundation, virtualization technology (VT) becomes a re-emerging technology. It is a decoupling technique that separates system software from hardware platform, while making applications to run pervasively. Many academic institutes and research labs from industries have devoted great efforts in various related aspects. In 2007, Chinese Ministry of Science and Technology initialized a basic research project (the 973 project) with 6 universities and 2 research institutes, aimed at various topics related to VT-based computing systems, such as VT architecture design philosophy, VT design for a single computing system, VT design for multiple computing systems, user environment for VT, security, reliability and trust issues related to VT, as well as performance evaluation and benchmarking for VT-based computing systems. In this talk, we will give insight for this project. Our experiences on VT research are discussed in detail, including power management, virtual machine live migration, memory/IO virtualization, and desktop virtualization.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, p. 41, 2010.

Recursive Queries Using Object Relational Mapping

Marta Burzańska[1], Krzysztof Stencel[1,2], Patrycja Suchomska[1],
Aneta Szumowska[1], and Piotr Wiśniewski[1]

[1] Faculty of Mathematics and Computer Science,
Nicolaus Copernicus University, Toruń, Poland
[2] Institute of Informatics, University of Warsaw, Warsaw, Poland
{quintria,stencel,papi,iriz,pikonrad}@mat.umk.pl

Abstract. Recent years witnessed continuous development of database query languages and object relational mappers. One of the research fields of interest are recursive queries. The first implementations of such queries for SQL has been introduced by Oracle in 1985. However, it was the introduction of recursive Common Table Expressions into the SQL:99 standard that made the research on this topic more popular. Currently most of the popular DBMS implements recursive queries, but there are no object relational mappers that support such queries. In this paper we propose extending existing ORMs with recursive CTE's support. A prototype of such an extension has been implemented in SQLObject mapper for the Python language. Tests have been conducted with PostgreSQL 8.4 database. Furthermore, recursive queries written using CTEs amount to be too complex and hard to comprehend. Our proposal overcomes this problem by pushing the formulation of recursive queries to a higher abstraction level, which makes them significantly simpler to write and to read.

1 Introduction

Since the 1970's Codd's publication [11], the relational data model has become a main standard for permanent data storage. Despite many attempts to introduce new standards like object algebra, nothing suggests that leading role of relational database systems is at risk. Modern relational DBMSs offer extensions to the relational model that widely extend their functionality. The research on such extensions has been greatly influenced by the need to express special data like bill-of-material, corporate hierarchy or graph structures. Querying of those structures is especially interesting for the authors of this paper. Recent years have shown an increase in research on the recursive query processing. The paper [1] discusses current State-of-the-Art in the field of SQL recursive Common Table Expressions. Most of the research around the recursive queries is focused on efficiency. The papers [7,2,8] discuss different aspects of optimisation - from query rewriting, through special on-the-fly placement of indexes during execution, to modifiable execution plans.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 42–50, 2010.

The SQL's recursive queries, despite their expressiveness, suffer from a big disadvantage - their construction is often too complicated for an average database user. Also, if a developer uses an ORM in order to benefit from recursive queries they have to be written in pure SQL by hand and explicitly passed to the database. The contribution of this paper is thus twofold. First, we propose a solution to the problem of recursive querying in object-relational mappers. Second, we combine the recursive queries with the simplicity of programming in Python. User is only required to provide some basic information and the tool provided by the authors will generate a proper query and pass it for execution. The resulting query formulation is significantly more comprehendible than written directly in SQL.

Further reasons of considering these queries and method of their implementation in example systems is being covered in Section 2.

Nowadays, the market of database applications has been dominated by object languages and relational DBMS used as a back-end. Those two different approaches combined result in different problems called the object-relational impedance mismatch. One attempt to solve those problems are object-relational mapping tools [9,10]. They will be discussed in section 3. In Section 4 we remind basic information on the technologies employed: Python and SQLObject. Section 5 presents the proposed solution to the problem of recursive queries in object-relational mappers. Section 6 concludes.

2 Two Motivating Examples

Let us consider the following situation: A standard example of a relation describing hierarchical structure is a corporate hierarchy representation. This relation joins employees with their supervisors. Let us assume we have a following relation `empl`:

Example 1. Relation with employees

```
 id | first_name | last_name | boss_id
----+------------+-----------+----------
  1 | John       | Travolta  |
  2 | Bruce      | Willis    |        1
  3 | Marilyn    | Monroe    |
  4 | Angelina   | Jolie     |        3
  5 | Brad       | Pitt      |        4
  6 | Hugh       | Grant     |        4
  7 | Colin      | Firth     |        3
  8 | Keira      | Knightley |        6
  9 | Sean       | Connery   |        2
 10 | Pierce     | Brosnan   |        3
 11 | Roger      | Moore     |        9
 12 | Timothy    | Dalton    |        9
 13 | Daniel     | Craig     |        1
 14 | George     | Lazenby   |         5
 15 | Gerard     | Butler    |         6
```

We would like to obtain information about all Travolta's direct and indirect subordinates. To acquire such data first we ought to find Travolta in the database. Then, we should search for all his subordinates, next their subordinates, etc. Using basic SQL:92 constructions such query would take the form of a union of series of queries:

```
Select id, first_name, last_name, boss_id
    From empl
    WHERE last_name = 'Travolta'
        UNION
Select e.id, e.first_name, e.last_name, e.boss_id
    From empl e JOIN empl c ON (e.boss_id = c.boss_id)
    WHERE c.last_name = 'Travolta'
        UNION
Select e.id, e.first_name, e.last_name, e.boss_id
    From empl e JOIN empl c_l1 ON (e.boss_id = c_l1.boss_id)
    JOIN empl c ON (c_l1.boss_id = c.boss_id)
    WHERE c.last_name = 'Travolta'
        UNION
...
```

Such construction has an obvious drawback - we have to know the depth of this hierarchy. Without this knowledge it is an impossible task. Fortunately the SQL:99 standard provides means of expressing the request for subordinates without this limitation. Those means are recursive views and recursive Common Table Expressions. The following example presents a query searching for Travolta's subordinates with the help of recursive CTE:

Example 2. Information about Travolta and all his direct and indirect subordinates.

```
WITH RECURSIVE rec_empl AS (
    SELECT id, first_name, last_name, boss_id
        FROM empl
        WHERE last_name = 'Travolta'
  UNION
    SELECT e.id, e.first_name, e.last_name, e.boss_id
        From empl e, rec_empl r
        WHERE e.boss_id = r.id
)
SELECT * FROM rec_empl
```

Example 1 presents hierarchical data which does not contain cycles.

Another example is going to consider a situation including natural cycles. Let us now consider the following relation describing connections between the cities:

```
 id | city_start | city_end | travel_time
----+------------+----------+-------------
  1 | Torun      | Warsaw   |        2.45
  2 | Warsaw     | Krakow   |         4.3
  3 | Torun      | Gdansk   |         2.2
  4 | Torun      | Olsztyn  |         1.8
  5 | Warsaw     | Kutno    |         1.3
  6 | Kutno      | Warsaw   |         1.3
  7 | Olsztyn    | Torun    |         1.8
  8 | Gdansk     | Warsaw   |           4
  9 | Kutno      | Torun    |         1.2
 10 | Lodz       | Krakow   |         3.5
```

For these structures it is natural to ask about connections between two cities placing a restriction on a number of transport changes. For example, we may ask about a connection from Torun to Zakopane with three transport changes at most. Unlike the previous example, this query may be expressed both using recursive CTE and a union of three queries. The union construction is fairly simple but spacious, thus it will be omitted in this paper. The following query presents the described connections request using a recursive CTE:

Example 3. Connections from Torun to Zakopane with a list of transport change cities and total time of travel (excluding time of waiting).

```
with recursive rcte as
  ( SELECT conns.id, conns.city_end, conns.travel_time,
       conns.city_start, ’ ’ || text ( conns.city_end ) as
       concat_city_end, ’ ’ || text ( conns.id ) as concat_id,
       ’ ’ || text ( conns.city_start ) as concat_city_start,
       conns.city_start as constant_city_start,
       conns.travel_time as sum_of_travel_time, 1 as level
  FROM conns
  WHERE conns.city_start=’Torun’
      UNION
  SELECT conns.id, conns.city_end, conns.travel_time,
      conns.city_start, text ( rcte.concat_city_end ||’, ’||
      conns.city_end ) as concat_city_end, text ( rcte.concat_id
      ||’, ’|| conns.id ) as concat_id,
      text ( rcte.concat_city_start ||’, ’|| conns.city_start )
      as concat_city_start, rcte.constant_city_start,
      rcte.sum_of_travel_time + conns.travel_time as
      sum_of_travel_time, rcte.level + 1 as level
  FROM conns, rcte
  WHERE rcte.city_end = conns.city_start AND rcte.level < 4
  AND rcte.concat_city_start NOT LIKE ’% ’||conns.city_start||’%’)
select * from rcte
  WHERE rcte.city_end=’Zakopane’
```

The examples given above present recursive queries expressed using PostgreSQL dialect [6]. For other database systems the notation may be slightly different [1]. We may notice from Example 3 that the recursive SQL queries may easily become difficult for a developer to understand and maintain. Also, writing such complicated queries in different application is a major inconvenience. This is why we propose a solution that allows the developer to focus on the essence of the queried data, while leaving the task of generating an appropriate query to the system.

3 Object-Relational Mappers

Modern database applications are mostly implemented using object-oriented languages. At the same time, the back-end for such applications are relational database systems. This creates a problem of translating information stored in objects into relational tuples that can be stored inside a relational database. This problem, more generally known as the impedance mismatch, has been thoroughly analysed for different approaches [9,10] and will not be discussed in this paper. Instead, we shall focus on one of its aspects - a method of expressing SQL queries inside the language used to build the application. This issue becomes even more troublesome when we permit connections to various databases using various SQL dialects. The tool used to bridge the gap between relational SQL and object-oriented programming language should be aware of differences among dialects, yet they should be irrelevant to the application developer using such tool. For example, there are many different JDBC drivers supporting various DBMS. Because of the variety of dialects, a JDBC driver should implement a method "jdbcCompliant()" [3] that tests the driver for compliance with JDBC API and SQL-92 standard. This means that each driver ought to restrict the usage of SQL constructions to only those supported by SQL-92 standard. Bearing in mind the rapid development of databases and programming languages in the last 15 years, this approach should be updated and SQL-92 should be replaced in this matter with a newer standard.

The ORM developers decided on a different approach. SQL queries are automatically generated according to the specific dialect's syntax dependent on the chosen DBMS. The mapper should have a set of rules describing how classes should correspond to relations, fields to relevant attributes, instances to tuples and references to foreign keys. Each supported SQL dialect should be described this way. The end application developer does not need to be aware of those rules and the data manipulation mechanisms used by the developer should be independent on the choice of the DBMS. Such solutions are possible for general application because every modern relational DBMS is based on the same data model and data access mechanism. Differences between DBMSs include their approach to specific optimisation methods, scalability and supported extensions. However, these aspects are not important in context of this paper.

4 SQLObject and Python

One of the choices to be made by this paper's authors was to select a base language for the experimental implementations. The choice was made based on the availability of ORMs, firm design and popularity of a language. Having considered many different languages the authors decided to use Python language. The main reasons behind this choice are the exceptional ease of prototype development and access to the source codes for ORM implementations. Also, Python's biggest drawback - its performance - has no significant impact on this work.

While choosing the base language the authors also have considered different ORMs. The prototype implementation could have resulted in alteration of the main ORM source code, thus the authors have decided to focus on open-source mappers. Among the Python's object relational mappers the SQLObject ORM has been chosen as the most developer-friendly. We will now briefly describe this ORM.

In SQLObject the configuration for specific data is set up using inheritance from a special class simply called the SQLOBject. Example 4 presents a configuration of an SQLObject mapping.

Example 4. Classes for `empl`, and `conns` tables will be formed respectively:

```
class Conns(SQLObject):
    cityStart = StringCol()
    cityEnd = StringCol()
    travelTime = FloatCol()


class Empl(SQLObject):
    firstName = StringCol()
    lastName = StringCol()
    boss = ForeignKey('Empl', default = None)
```

Example 5. Employees with their last name "Craig" and the last names of their bosses:

```
craigs = Empl.select(Empl.q.lastName == "Craig")
for e in craigs:
    print e.boss.lastName
```

Example 5 uses so called `q`-magic. It is an SQLObject's technique that allows for defining search conditions in Python language. Such condition is automatically mapped to SQL when needed. The `Empl.q.lastName` object is a sub-object of `Empl` class. It represents the information about the `last_name` attribute of the `empl` database relation. More detailed information about the `q`-magic technique can be found at [5].

5 Recursive Query in SQLObject

The main idea of the proposed recursive query support is to extend the mechanisms provided by ORMs with additional methods for passing recursive queries.

Those methods should be able to create an SQL query representing a recursive CTE in a specific dialect of the target database. From the users point of view they should represent a set of unmodifiable data - a recursive view. Having those considerations in mind the authors have decided to base their designs on the standard `SQLObjectView` class. Instances of this class can be queried, but cannot be altered. The proposed user interface class is `ViewRecursive` class. In order to create a recursive query, a developer should declare a new class inheriting from `ViewRecursive` and equip it with a set of class attributes described below:

- `tables = []` - a list of SQLObject classes representing tables that will be used to form the recursive query.
- `recursiveOn` - a field of the base table used to join it with a CTE in a recursive step.
- `recursiveTo` - a field of the CTE used to join in with the base table in a recursive step.
- `initialCondition = []` - a list of conditions that should be met by tuples from the initial SELECT subquery of the recursive query.
- `predicateList = []` - a list of conditions that should be met by resulting tuples (instances).

The fields `recursiveOn` and `recursiveTo` form an equality predicate in a recursive part of the CTE's definition. This predicate is the basis of the recursion. The `predicateList` represents predicates used in the outer SELECT query that uses the CTE. The following attributes are optional - they are used to provide additional functionality to the data search:

- `maxLevel` - upper boundary on recursion depth.
- `summands = []` - a list of fields, which are going to be summed up. It should be noted here, that the mapper does not check type compatibility.
- `concats = []` - a list of fields, which results will be concatenated with a single whitespace character as a space between arguments.
- `constants = []` - a list of fields, which values are going to be repeated in all iterations.

The `maxLevel` attribute is especially useful in case of searching through a cyclic structure. Lack of this attribute definition in such case may lead to infinite loop calculations. The usage of `constants` is convenient in situations when during the recursion steps the tree invariants are being created. T he tree invariant is an attribute value that once generated in an initial step is constant for all tuples generated out of an initial tuple. The existence of the tree invariant allows for application of additional optimisation transformations described in [2].

Examples 6 and 7 present the usage of this construction to express recursive queries described earlier.

Example 6. A definition of a class corresponding to Example 2:

```
class Subordinates(ViewRecursive):
    tables = [Empl]
    recursiveOn = Empl.q.boss
    recursiveTo = Empl
    initialConditions = [Empl.q.lastName == "Travolta"]
```

This class has a method `getRecursive` inherited from `ViewRecursive`. This method generates a query presented in Example 2 using the information provided by the attributes of the class. Next, it creates read-only objects containing fields: `id`, `first_name`, `last_name`, `boss_id`. The `boss_id` field is a pointer to proper `Empl` class instance. The declaration of this class is simpler than the query from Example 2, but the gain is not too striking yet.

Example 7. A definition of a class corresponding to Example 3.

```
class Travel(ViewRecursive):
    tables = [Conns]
    recursiveOn = Conns.q.cityStart
    recursiveTo = Conns.q.cityEnd
    maxLevel = 4
    summands = [Conns.q.travelTime]
    concats = [Conns.q.cityEnd, Conns]
    constants = [Conns.q.cityStart]
    initialConditions = [Conns.q.cityStart == "Torun"]
    predicateList = [Conns.q.cityEnd == "Zakopane"]
```

Method `getRecursive` of the `Travel` class generates a query from Example 3. This query results in a collection of objects with the fields: `conns`, `city_end`, `travel_time`, `city_start`, `concat_city_end`, `concat_id`, `concat_city_start`, `constant_city_start`, `sum_of_travel_time`, `level`. The `constant_city_start` is the tree invariant described above. Now, the simplicity of recursive queries written according to our proposal is apparent. It is enough to compare the complex code from Example 3 with easily comprehendalble class definition from Example 7.

To acquire the requested data from the database the `getRecursive` method is called. This function first generates a recursive query, which is then passed to the database. The result of the query is used to create a collection of read-only objects representing the data. I n particular, calling `Travel.getRecursive()` will generate and execute query from Example 3. Although the `Travel` class instances are read-only, they contain pointers to fully modifiable objects. This correlation is presented by Example 8:

Example 8. `emps = Subordinates.getRecursive()`

```
for e in emps:
    if e.lastName == "Craig":
        e.boss.firstName = "Bruce"
```

In this query the variable `e` representing Craig is read-only. However, it points to the `Bruce Willis` object, which is the `Empl` class instance. This class of objects allows for data modifications.

The proposed solution significantly shortens the time required for preparing a recursive query. It also has a great impact on readability and maintainability of the code. A developer no longer needs to focus on the syntax and complexities of recursive SQL queries. Instead, he/she may concentrate on the pure data he/she wants to gather. Additional benefit of our method is that it would provide a cross-platform solutions with only small adjustments in configuration to fit specific dialects.

6 Conclusions and Future Work

In this paper we presented a proposition how to incorporate recursive queries into object-relational mappers. So far, such mappers do not facilitate recursive queries. Furthermore, the presented method allows expressing recursive queries in a noteworthy simpler way than it is now possible in SQL.

Future research plans include developing prototypes for other databases, in particular Oracle, IBM DB2 and MS SQL Server. The next step would be porting the presented algorithm to django models mapper. Another interesting work would be translation of the proposed solutions to other ORMs and languages, for example Java and C#. In particular the problem of integrating recursive queries into LINQ for C# seems a very promising research topic. Note also, that our method do not uses strings to express any metainformation in the query. This means that in statically typed languages queries can be type checked in every single detail.

References

1. Boniewicz, A., Burzanska, M., Przymus, P., Stencel, K.: Recursive query facilities in relational databases: a survey. manuscript sent for DTA (2010) [if both papers are accepted, the editor may change this reference]
2. Burzanska, M., Stencel, K., Wisniewski, P.: Pushing Predicates into Recursive SQL Common Table Expressions. In: Grundspenkis, J., Morzy, T., Vossen, G. (eds.) ADBIS 2009. LNCS, vol. 5739, pp. 194–205. Springer, Heidelberg (2009)
3. JDBC Driver class documentation, http://java.sun.com/j2se/1.5.0/docs/api/java/sql/Driver.html
4. SQLObject, http://www.sqlobject.org/
5. q-magic, http://www.sqlobject.org/SQLObject.html#q-magic
6. Recursive queries in PostgreSQL, http://www.postgresql.org/docs/8.4/static/queries-with.html
7. Ghazal, A., Crolotte, A., Seid, D.Y.: Recursive SQL Query Optimization with k-Iteration Lookahead. In: Bressan, S., Küng, J., Wagner, R. (eds.) DEXA 2006. LNCS, vol. 4080, pp. 348–357. Springer, Heidelberg (2006)
8. Ordonez, C.: Optimization of Linear Recursive Queries in SQL. IEEE Trans. Knowl. Data Eng., 264–277 (2010)
9. Melnik, S., Adya, A., Bernstein, P.A.: Compiling mappings to bridge applications and databases. In: ACM SIGMOD, pp. 461–472 (2007)
10. Keller, W.: Mapping objects to tables: A pattern language. In: EuroPLoP (2007)
11. Codd, E.F.: A relational model of data for large shared data banks. Communications of the ACM (1970)

Agent Based Ontology Driven Virtual Meeting Assistant

Phil Thompson, Anne James, and Emanuel Stanciu

Distributed Systems and Modeling Research Group
Coventry University
United Kingdom
{P.Thompson,A.James,E.Stanciu}@Coventry.ac.uk

Abstract. It is widely recognized that much management time and effort is expended in attending meetings and being such a costly resource it is important that the decision to attend a meeting is not taken lightly. Because of this many busy executives are rejecting face-to-face meetings in favour of virtual meetings involving both high and low complexity technological solutions. This paper proposes an agent based, ontology driven meeting assistant designed to support virtual meetings. Situated on a graph database and operating over a peer-to-peer network the ontology will provide the basic vocabulary for the meeting and the semantics required to achieve a dialogue between the agents and the human interface used by the meeting assistant.

Keywords: Virtual meetings, Peer-to-peer, Multi-agent systems, Ontology, Graph database.

1 Introduction

This paper develops ideas explored by the author in similar papers to provide an environment to support meetings [1] using agents recognizing that meetings are becoming increasingly virtual in nature. That is to say that the meetings take place with the participants in their home locations, “meeting” their fellow participants over a network using a computer. The increase in the use of virtual meetings [2] has come about because high ranking company executives are involved more and more in meetings with global partners involving travel over large distances with the ensuing high costs [3]. It is costly not only from the standpoint of the actual air travel, accommodation and other expenses which are involved, but also because of the amount of the executives’ time which is spent on route to and from meetings [4]. Even though virtual meetings are not yet considered to be more effective than face-to-face meetings one writer [5] cites the Wall Street Journal study which reported that out of 2300 business leaders 60% of the respondents used virtual facilities “very frequently”. These virtual meetings are becoming more common and are being used by companies for prestigious occasions like sales conferences as well as business meetings. There are many ways for virtual meetings to be supported. These range from the simple, involving PC’s and the use of software such as Microsoft Office Live Meeting, together with webcams and microphones over a broadband or similar link, to the more sophisticated

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 51–62, 2010.

Telepresence [6,7] which, through the use of large screens gives the impression of all the meeting attendees being present in the same room, even though they are separated by thousands of miles. Companies have also taken advantage of Second Life, meeting other people in a virtual world disguised in the form of an avatar. [8].

It has long been recognized that in order to manage a meeting effectively a good facilitator is required [9,10,11,12]. The role of the facilitator includes: ensuring the objectives of the meeting are met and that the meeting is not going off track; allowing all the meeting participants to have input to the meeting; preventing some individuals dominating; and summarising and clarifying discussion where necessary. Clearly, if a facilitator is needed in a face-to-face meeting where all the participants are in the same room, it is more necessary in a virtual environment. In a virtual environment, the ordered scheduling of contributions into the discussion becomes much more difficult. To prevent everybody talking at once and controlling the meeting requires a lot of concentration and skill by the facilitator. Also, the fact that virtual meetings tend to be asynchronous with participants leaving and re-joining the meeting as it runs its course presents other challenges [13] compared to face-to-face meetings. For these reasons to provide computer assistance to manage the participants and to take away some of that responsibility from the facilitator, allows that individual to focus more on the content of the discussion than the mechanics.

The use of computer based agents to assist humans in the performance of tasks is a dream [14] that is still yet to achieve a "killer application" but promises to become the means by which computers can perform less important tasks while enabling humans to concentrate on more important ones.

In order to assist in the facilitation of meeting an intelligent agent requires a knowledge of the structure and vocabulary of a meeting. An ontology provides the vocabulary of words and the relationship between them, providing meaning, enabling computers to make sense of a particular domain. A graph database is a means by which ontologies can be created and maintained making it an ideal vehicle for the intelligent agent to "understand" its environment.

In order to address some of the problems mentioned above these now available technologies have been utilized in the development of a prototype "meeting assistant" designed to support the facilitator of the virtual meeting environment. This paper describes the prototype. Section 2 identifies the objectives and theoretical foundation behind the design, Section 3 gives a functional description, Section 4 describes the technical design, Section 5 explains the way the prototype has been implemented and Section 6 provides conclusions and outlines our future work.

2 Objectives and Theoretical Foundation

The objective of this paper is to report on the work to use a multi-agent approach as a possible solution to the problem of managing virtual meetings. By using an ontology to provide the means for agents to make sense of and give meaning to their environment, they can be used to assist the meeting facilitator in performing that meeting management role. This work draws on the theoretical foundation set by other researchers in these fields and attempts to introduce some new ideas to further this work. This section explains how these ideas have built on this theoretical foundation giving a brief introduction to multi-agent systems, ontologies and peer-to-peer networks.

2.1 Multi-Agent Systems

Multi-agent systems have been described as “computational systems in which several artificial "agents", which are programs, interact or work together over a communications network to perform some set of tasks jointly or to satisfy some set of goals” [15]. They are also described as either “benevolent or self-interested” in pursuit of their goals and autonomous in their choices. Although a truly independent agent should be able to choose whether or not they complete a task they would not be much use in a situation where the task is important. The particular agents that are described in this paper are benevolent rather than self-interested sharing the goals of their master (being the facilitator or the participants of the meeting) according to the requests made of them. They pursue those goals with the set of skills that they have been programmed to use. This means that they release some of their autonomy to co-operate with their master.

The Open Multi-Agent System (OMAS) [16] provides the infrastructure for the application. OMAS uses a concept of personal assistants which provides a human interface to allow a mixed environment of humans and agents in the design of applications. It provides skeleton agents to which can be added LISP code which gives the agents the “skills” required to perform the actions required by the applications. The mechanisms to allow natural language statements to be passed to agents from the personal assistant are included and these statements are then parsed to extract the “performatives” and associated information. The performatives are verbs which require the agent to execute a particular action according to its implemented skills. The creation of the dialogue and vocabulary which is necessary to make sense of the statements must also be implemented in LISP code although again skeleton procedures are provided to simplify this process. The human interface of the personal assistant provided by OMAS is suitable for prototyping the processing of these statements but what is required for actual applications is a procedure more specific to the needs of that application.

The meeting assistant described in this paper has provided an interface which is designed specifically for a virtual meeting environment and attempts to make the process of creating vocabulary and dialogue more configurable by the application designer. This is achieved by using a database which can be created and maintained without recourse to LISP code.

2.2 Ontologies

Ontologies have their origins in philosophy but work started in the 1990’s to use them to model information systems [17]. A definition given at that time was, “a specification of conceptualization”. This was further developed into, “a specification of a representational vocabulary for a shared domain of discourse – definitions of classes, relations, functions and other objects”. This has evolved today to a comprehensive specification of an ontology in the form of the Web Ontology language (OWL), Resource Description Framework (RDF) and Vocabulary Description Language (RDFS) which have become the adopted standards for applications embodying ontologies into their design. These are the standards which have been built into the ontology used in the Meeting Asssistant described in this paper. The ontology has been implemented using the ALLEGROGRAPH graph database. The ontology database has then been used to give the agents a vocabulary for the meeting application concepts and objects and the relationships between them, to allow them to make sense of that domain. The idea of using an ontology in an application using multi-agents is not new and is a concept employed by OMAS described above. The ontology in OMAS is implemented using the LISP language which although made simpler by the use of skeleton procedures requires a working knowledge of the language to configure the information. What is different in the application described in this paper is that the ontology is embedded into a database and uses RDF statements as the means of describing the word vocabulary, object properties, relationships and command list of the application making configuration easier. The command-list on the ontology will be used to validate the keywords addressed to agents from the meeting participants.

2.3 Peer-to-Peer

In a peer-to-peer network [18] no central server is required. Each peer brings enough extra resource to support its own membership of the network. Peer-to-peer computing also allows shared information to be, "accessible by other peers directly, without passing intermediary entities. When a new member joins the team information can be sourced from any of the other members. Although this creates data redundancy it does offer data resilience by providing an answer to the single point of failure problem encountered in central server systems because no central repository is used for information. The scalability offered by peer-to-peer networks means that it is easy for new participants to join the meeting. In a conventional client-server network extra resources have to be provided at the server as more and more nodes join the network.

Peer-to-peer is used by the application in this paper to broadcast messages so that any contribution for a participant in the meeting is sent to all the participants to maintain their record of the conversation. It also allows participants to be able to communicate with each other without any central server being involved. OMAS uses peer-to-peer as the means of sending and receiving messages from one computer to another. The scalability of a peer-to-peer network allows new meeting participants to join the meeting with minimum impact on the rest of the system.

3 Functional Description of Prototype

To support the typical virtual meeting environments referred to above there needs to be both control over the dialogue between participants as well as the need to assist the facilitator in managing the meeting. To control the dialogue what is required is an orderly transfer of the dialogue from one participant to the next. Ideally it should not be possible for more than one person to contribute to the meeting at the same time and the time allowed, to give opportunity to all the participants to make a contribution should be carefully metered. The work of the facilitator is to be concentrated on bringing participants in to the discussion in turn, allowing the person who is the most knowledgeable on an issue to lead. Any offline tasks which would interfere with this process should be handled by the meeting assistant. These tasks would include such peripheral activities as issuing agendas or discussion papers to new arrivals at the meeting. This becomes especially important because the nature of the virtual meeting is asynchronous, with participants joining and leaving the meeting more so than happens in a synchronous or face-to-face meeting. The usual expectation in the face-to-face meeting is that people arrive at the beginning, receive all the supporting documentation, take part in the meeting and then leave at the end. A virtual meeting is likely to last longer, sometimes spread over many days and people would be more selective over the agenda items to which they contributed.

Using a suitable input device, the participant will, on registering for a meeting, be able to access any information relating to previous related meetings including a record of the discussion, summary of agenda items and actions given to individuals together with any supporting documentation. On requesting to join the meeting the participant will receive a list of all the meeting participants together with the agenda for the meeting and a summary of the meeting so far if the meeting was in progress. The participant will have a visual display of the typed contributions of all the other meeting attendees as they occur and will be able to make a contribution if required. All requests to contribute to the meeting will be directed to the facilitator who will cue the participant when it is their turn. The participant will inform the facilitator if they wish to leave the meeting which could occur through interrupts at their location or simply because they have no interest in the particular agenda item. This will remove that individual from the list of current attendees.

The facilitator will move through the agenda items in turn and after enough discussion has taken place will summarise the main decisions, record any actions against individuals together with the date the action should be completed and then close the agenda item. If it is judged that there has not been enough discussion on an item or because information is needed from somebody not present at the meeting the facilitator may choose to leave the agenda item open and proceed with the meeting. When all the agenda items have been discussed the facilitator will either suspend or close the meeting depending on whether any agenda items are still open. The facilitator may choose to interrupt the meeting or the contribution of a participant at any time. If the meeting is interrupted in this way all participants will be notified of the reason and informed when the meeting will continue. This could happen if the facilitator needed a break in the proceedings or wished to suspend an agenda item to bring a more important item forward for discussion or schedule in a previous agenda item left open because a key individual has now joined the meeting. Another option open to the facilitator particularly if there is a need to attend to some other business, would be to pass the facilitator responsibility for the meeting to another individual. In practice any one of the meeting participants could take control of the meeting simply by giving them access to the functions normally given to the facilitator. A third option would be to allow the meeting to run on "automatic" at least as long as there are still requests to contribute from participants when the meeting would be temporarily suspended until the return of the facilitator. This would mean that there would be no supervision of the dialogue with the atttendant risk of the discussion moving "off topic" and open to abuse.

When agenda items require formal agreement by the meeting participants they will be able to propose or second a motion and then agree, disagree or abstain in any voting activity. This will be recorded against the agenda item for future reference. Voting preferences will not be recorded against any individual in the meeting apart from indicating whether they have voted or not. This will allow votes to be cast by individuals later before the agenda item is closed.

The ontology will contain a vocabulary of the terminology used in the meeting and any relationships between words or concepts within that vocabulary. So for example the ontology will "understand" that the "agenda-item" relates to an "agenda" which itself relates to a "meeting" and so on. Also contained in the ontology will be a series of "performatives" which will allow the participants to issue natural language statements to the meeting assistant which will be recognized by matching against the ontology and result in certain actions being performed by the assistant. For example the use of the performative "get" together with the word "agenda" will result in the meeting agenda being sent to the participant. Properties will be stored against these terms to help the meeting assistant decide the action required. For example against the "agenda-item" will be a filename or URL to allow the information to be accessed from a database or on the web. The ontology will be stored on a database to allow easy updating of the terminology, relationships and properties for different types of meeting.

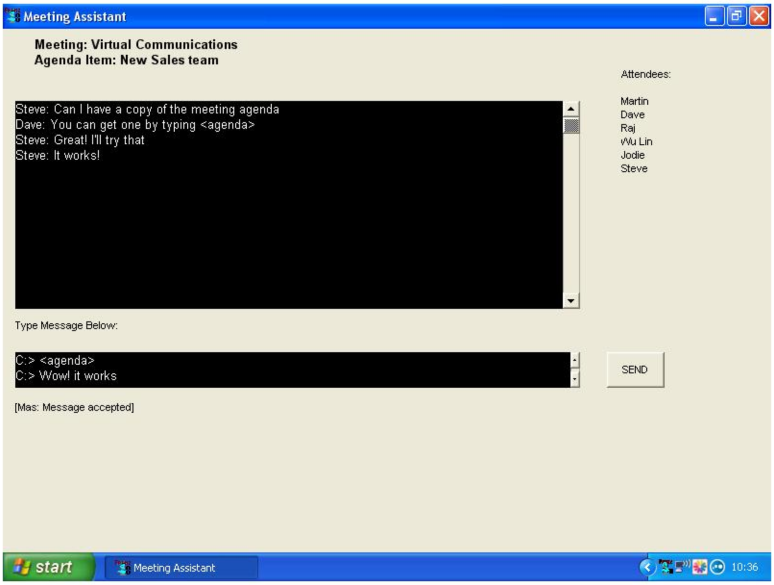
For efficient management of the meeting there needs to be a facilitator. The facilitator needs to create an agenda where the agenda item sequence is clear for all participants and an estimated time for discussion for each item should be added. Ideally somebody should be asked to lead the discussion on each agenda item. At the end of each agenda item the facilitator should summarise findings and then clearly identify what actions are required and who is to perform the action [14]. A date should be associated with action to ensure feedback.

The participants on joining the meeting should be given a copy of the agenda and any other supporting documents. The facilitator should control who speaks and when they speak and should be able to interrupt if required to seek clarification. Participants should be able to signal to the facilitator to gain access to the meeting; when they wish to speak; if they do not understand; if they want to move progress; if they wish to leave the meeting, etc. They should be able to terminate a session and resume at a later time in case they do not need to be present for certain agenda items. At the end of the meeting they should be provided with a copy of the action list by the facilitator.

4 Technical Design of Prototype

The prototype environment we designed consists of control windows for the participant and facilitator. These control windows are used to display the discussion and allow communication and also provide access to the infrastructure which supports the meeting. The environment is designed to allow all communication between participants, facilitator and the meeting assistant agent to be controlled from the one window using a conversational dialogue.

The Input-Output window in Figure 1 is the first window to appear after login. It gives the participant a list of all the other meeting attendees by displaying their names down the right hand side of the screen. The discussion itself appears in the main part of the window with a scrolling record of the contributions made showing the name of the participant together with their typed contribution. Participants wishing to contribute to the meeting discussion will type their contribution and then hit the "send" button which will display the entry in the scrolling record after the entry of the current participant has been completed.



Meeting Assistant

Meeting: Virtual Communications
Agenda Item: New Sales team

```
Steve: Can I have a copy of the meeting agenda
Dave: You can get one by typing <agenda>
Steve: Great! I'll try that
Steve: It works!
```

Attendees:

Martin
Dave
Raj
Wu Lin
Jodie
Steve

Type Message Below:

```
C:> <agenda>
C:> Wow! it works
```

[Mas: Message accepted]

Fig. 1. Input-Output Window

The participant can also communicate to others in the meeting by highlighting them before hitting the send button which will direct the message to those selected. This facility can be allowed or disallowed at the discretion of the meeting facilitator. It could be useful to allow the facility if discussion is required between members before a decision can be taken. Messages received back will only appear on the screens of the selected participants and the original sender.

Participants will obtain information about the meeting by the use of keywords they type into the message. These messages will be targeted at the meeting assistant which will process them accordingly. The list of keywords and how to use them will be sent to the participant on successful login. They will include: get agenda, get agenda item, list attendees, etc. Any information sent in response to these messages will be displayed in pop-up windows on the requestor's main screen.

The facilitator (who could be at the same time a meeting participant) will also use the display and communication window. Using that window, agenda items will be introduced; the flow of the meeting will be interrupted when required; the agreed actions will be typed when all the discussion on an agenda item has been completed and the meeting will be closed. All this will be visible to the other meeting participants in the scrolling display, which will become a permanent record of the meeting.

All the messages from participants will be directed to the meeting assistant which is an intelligent agent who through a system of keywords embedded in the messages will be able to perform the necessary processing of the message. Keywords identifying the sender and target of a message will precede and follow the message and the agent will direct the message accordingly, messages being displayed on the scrolling record in the display and communication window of the target participant. Normal messages not containing a target keyword will be directed to all participants.

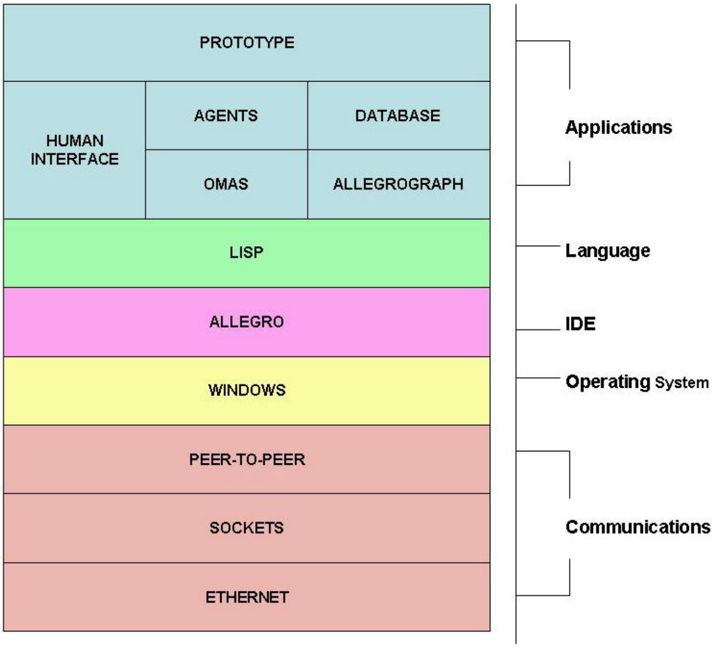


Fig. 2. Prototype Implementation Layers

Where the message target keyword specifically identifies the meeting assistant this will indicate the requestor requires information about the meeting which will necessitate sending messages to other agents called meeting agents who will process the message and obtain the required information. The meeting agents will use the other keywords in the message to access the meeting ontology and extract the required information to be sent back to the participants.

5 Implementation of Prototype

The implementation layers of the prototype are shown in Figure 2. The application layer of the prototype is made up of the human interface, the agents and the database. The human interface is where the human participants and facilitator interact with the system. The agents, consisting of the meeting assistant and meeting agents operate within the OMAS environment and interact with the human interface by sending and receiving messages. The graph database using the ALLEGROGRAPH package is updated by the meeting agents and holds the meeting structure, vocabulary and list of commands in the form of an ontology.

The applications have been developed in the LISP language using ALLEGRO as the Interactive Development Environment (IDE) and operate under the control of the WINDOWS operating system.

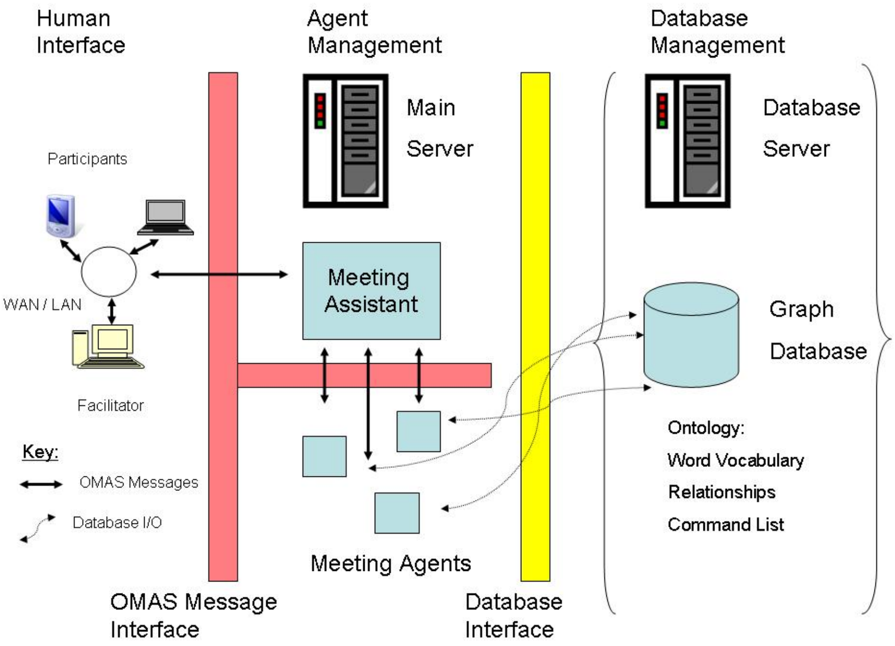


Fig. 3. The Protoype Design

The communications layer controls the interaction between the different nodes on the Ethernet network. The facilitator and the participants each occupy a node on the peer-to-peer network. The communications between nodes is implemented by the use of sockets which identify the IP address and port number of each node. The same mechanism is used to communicate with the agent infrastructure through the use of messages being sent via the sockets to and from the human interface and the agent infrastructure.

The OMAS agents process the messages according to their implemented “skills” in order to service the requests coming from both the facilitator and the participants. The prototype is shown diagrammatically in Figure 3 which also illustrates how the components of the virtual meeting environment can be located on different computers.

The meeting ontology is achieved through the use of a graph database which, through the use of a triples datastore, identifies the terminology used in the meeting environment, the relationship between objects reflected in those words and the properties of the objects. An extract of the ontology for the meeting is shown in graph form in Figure 4, and in the Resource Description Framework (RDF) in Figure 5 which illustrates how this information is held in the database in triples format. The equivalent graph is shown for the Command List in Figure 6.

When the meeting assistant recognizes a command keyword in the input line from the participant the command is extracted and sent to one of the meeting agents. The meeting agent first validates the command performative against the graph database returning an error if it does not match the database. If the performative is valid the agent processes the rest of the command and by breaking it up into recognizable

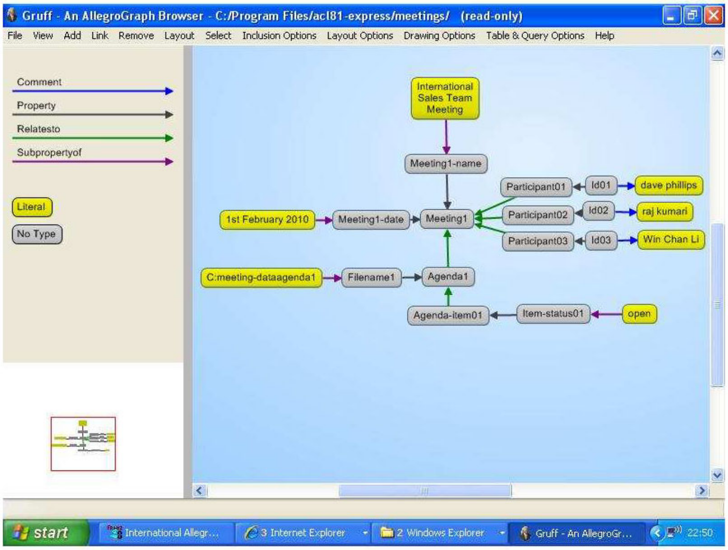
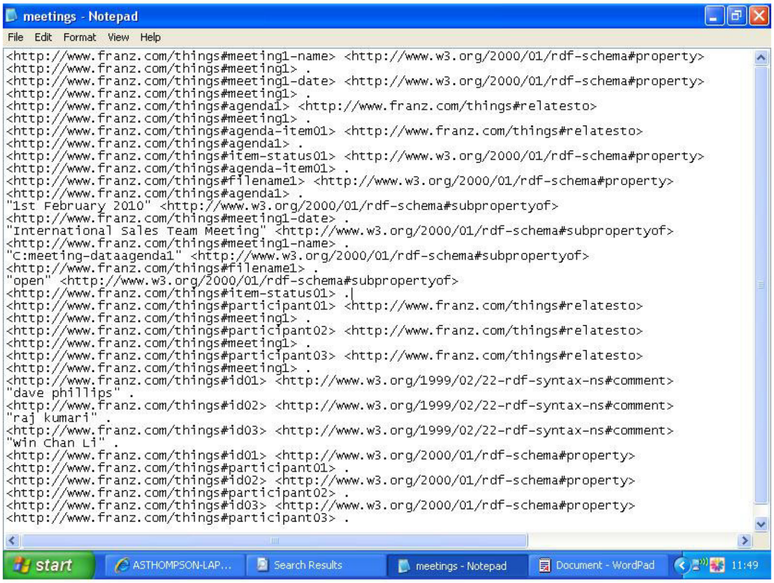


Fig. 4. Meeting Ontology Graph



```
<http://www.franz.com/things#meeting1-name> <http://www.w3.org/2000/01/rdf-schema#property>
<http://www.franz.com/things#meeting1> .
<http://www.franz.com/things#meeting1-date> <http://www.w3.org/2000/01/rdf-schema#property>
<http://www.franz.com/things#meeting1> .
<http://www.franz.com/things#agenda1> <http://www.franz.com/things#relatesto>
<http://www.franz.com/things#meeting1> .
<http://www.franz.com/things#agenda-item01> <http://www.franz.com/things#relatesto>
<http://www.franz.com/things#agenda1> .
<http://www.franz.com/things#item-status01> <http://www.w3.org/2000/01/rdf-schema#property>
<http://www.franz.com/things#agenda-item01> .
<http://www.franz.com/things#filename1> <http://www.w3.org/2000/01/rdf-schema#property>
<http://www.franz.com/things#agenda1> .
"1st February 2010" <http://www.w3.org/2000/01/rdf-schema#subpropertyof>
<http://www.franz.com/things#meeting1-date> .
"International Sales Team Meeting" <http://www.w3.org/2000/01/rdf-schema#subpropertyof>
<http://www.franz.com/things#meeting1-name> .
"C:meeting-dataagenda1" <http://www.w3.org/2000/01/rdf-schema#subpropertyof>
<http://www.franz.com/things#filename1> .
"open" <http://www.w3.org/2000/01/rdf-schema#subpropertyof>
<http://www.franz.com/things#item-status01> .
<http://www.franz.com/things#participant01> <http://www.franz.com/things#relatesto>
<http://www.franz.com/things#meeting1> .
<http://www.franz.com/things#participant02> <http://www.franz.com/things#relatesto>
<http://www.franz.com/things#meeting1> .
<http://www.franz.com/things#participant03> <http://www.franz.com/things#relatesto>
<http://www.franz.com/things#meeting1> .
<http://www.franz.com/things#id01> <http://www.w3.org/1999/02/22-rdf-syntax-ns#comment>
"dave phillips" .
<http://www.franz.com/things#id02> <http://www.w3.org/1999/02/22-rdf-syntax-ns#comment>
"raj kumari" .
<http://www.franz.com/things#id03> <http://www.w3.org/1999/02/22-rdf-syntax-ns#comment>
"Win Chan Li" .
<http://www.franz.com/things#id01> <http://www.w3.org/2000/01/rdf-schema#property>
<http://www.franz.com/things#participant01> .
<http://www.franz.com/things#id02> <http://www.w3.org/2000/01/rdf-schema#property>
<http://www.franz.com/things#participant02> .
<http://www.franz.com/things#id03> <http://www.w3.org/2000/01/rdf-schema#property>
<http://www.franz.com/things#participant03> .
```

Fig. 5. Meeting Ontology Graph – Vocabulary

elements uses those elements to perform the processing. The result of the command could update the database with information passed from the participant or extract information from the database to be sent back to the participant. The processing of the database is performed by building up LISP queries from code fragments held on the database according to the type of processing required.

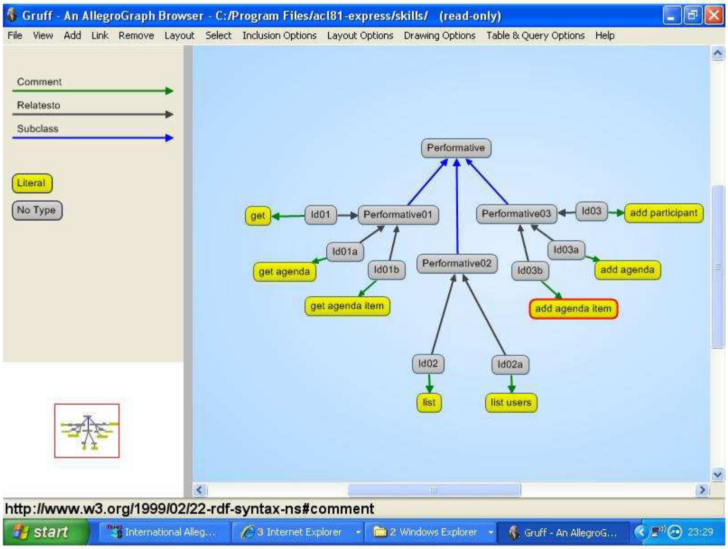


Fig. 6. Meeting Ontology Graph - Commands

6 Conclusions and Future Work

Techniques and solutions have been referred to in this paper which use computing technology for supporting virtual meetings and this is now an established option for business. However, the novelty of the system described in this paper is that the natural language dialogue from the meeting participants and facilitator is interpreted by the use of an ontology and given meaning to allow requests to be performed by "intelligent agents" in response. It provides the means to manage virtual meetings by allowing the participants to see the contributions of the other participants, add their own contributions and also to obtain supporting information and manage the agenda through a simple natural language interface. The facilitator can concentrate on the important task of co-ordinating the agenda and the contributions of participants while being assisted by multiple computerized agents with the less important tasks. The structure and vocabulary of the meeting understood by the agents is implemented as an ontology which, residing on a database can be easily created and maintained.

Future work will be concentrated on further refinement, trialling the application in a real meeting situation, analysing the results and reporting the outcomes.

References

1. Thompson, P., Iqbal, R.: Supporting the Social Dynamics of Meetings using Peer-to-peer Architecture. In: Proceedings of the 15th International workshops on Conceptual Structure, CSCWD 2007, pp. 170–176. Springer, Heidelberg (2007)

2. Thompson, P., Iqbal, R., James, A.: Supporting collaborative virtual meetings using multi-agent systems, cscwd. In: Proceedings of 13th International Conference on Computer Supported Cooperative Work in Design, pp. 276–281 (2009)
3. Armfield, R.: Virtual meetings save real money, Bank Technology News 23(7), 13 (2010) (AN52411030)
4. Ellis, C., Barthelmess, P.: The Neem Dram. In: Proceeding of the 23rd Conference on Diversity in Computing (TAPIA 2003), pp. 23–29. ACM Press, New York (2003)
5. Boehmer, J.: Harvard study shows face-to-face meeting value, rising virtual interest. Meeting News 33(12), 9, 1p, 1 Chart (2009)
6. Scofidio, B.: Why Should(n't) You Manage Virtual Meetings. Corporate Meetings & Incentives 27(8), 4, 1 (2008)
7. Bulkeley, W.: Better Virtual Meetings. Wall Street Journal-Eastern Edition 248(75), B1-B5 (1987)
8. Kharif, O.: The Virtual Meeting Room. Business Week Online, 6, 1 (2007) AN 24781227
9. Black, A.C.: Getting The Best From Virtual Meetings. Bloomsbury Business Library – Manage Meetings Positively, 60–71, 12 (2006)
10. Miranda, M., Bostrom, P.: Meeting facilitation: process versus content interventions. Journal of Management information systems 15(4), 89–114 (1999)
11. Barker, A.: Crash Course in having effective meetings. Management Today 22(2/3) (June 2008) (AN 32828858)
12. Scofidio, B.: People Management.: How to have effective meetings, vol. 14(20), p. 45, 1 (2008)
13. McQuaid, M., et al.: Tools for distributed facilitation. In: Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, p. 10 (2000)
14. Fisher, L.: Make way for intelligent agents, cited by: Strategy & Business. In: Booz & Company 2010 (1995)
15. Lesser, V.: Encyclopedia of Computer Science, 4th edn., pp. 1194–1196 (2003)
16. Barthes, J.-P.A.: OMAS – A flexible multi-agent environment for CSWD. In: Computer Supported Co-Operative Work in Design 2009, pp. 258–263 (2009)
17. Gruber, T.R.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5(2), 199–220 (1993)
18. Kellerer, W.: Dienstarchitekturen in der Telekommunikation – Evolution, Methoden und-Vergleich. Technical Report TUM-LKN-TR-9801 (1998)

Topic Number Estimation by Consensus Soft Clustering with NMF

Takeru Yokoi

Tokyo Metropolitan College of Industrial Technology, Japan
takeru@s.metro-cit.ac.jp

Abstract. We propose here a novel method to estimate the number of topics in a document set using consensus clustering based on Non-negative Matrix Factorization (NMF). It is useful to automatically estimate the number of topics from a document set since various approaches to extract topics can determine their number through heuristics. Consensus clustering makes it possible to obtain a consensus of multiple results of clustering so that robust clustering is achieved and the number of clusters is regarded as the optimized number. In this paper, we have proposed a novel consensus soft clustering algorithm based on NMF and estimated an optimized number of topics by searching through a robust classification of documents for the topics obtained.

Keywords: Consensus Clustering, Estimation of the number of topics, Soft Clustering, Topic extraction.

1 Introduction

Such a stream of text documents are available on the Internet that in order to organize them all, focus is placed on the topics and various approaches of topic extraction from the text documents have been proposed, e.g., clustering algorithm, matrix factorization and other statistical methods. Topic extraction by a clustering algorithm regards a centroid of each document cluster as a topic in a document set [1]. In all of the topic extraction approaches, it is important to address how to determine the number of topics where the number of topics equals the number of clusters for topic extraction by clustering. However, the number of topics in major topic extraction methods is arbitrarily determined through user experimentation. In fact, some approaches to determine the number of clusters such as the use of information criteria, i.e., the Akaike Information Criteria (AIC), the Bayesian Information Criteria (BIC), etc., also exist in the field of clustering [2][3]. In the field of topic extraction, assuming the semantic distribution, i.e., a dirichlet process [4], is proposed to automatically determine the number of topics.

The method using consensus clustering is proposed as one approach to determine the number of clusters. It explores the optimized number from the results of consensus clustering with the various numbers of clusters by empirical cumulative distribution (CDF) [5]. In recent years, an approach using Non-negative Matrix Factorization

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 63–73, 2010.

(NMF) as the main clustering algorithm of consensus clustering has been proposed [6]. In addition to using NMF as the main algorithm, an approach to determine the number of clusters has been also proposed [7] with a Cophenetic correlation coefficient (CPCC) [8]. NMF is appropriate for consensus clustering since the random initialization that the factorized matrices are randomly initialized is performed at every clustering repetition. In addition, this algorithm [7] can be efficiently applied for gene data identification.

NMF has also been applied to a document set and each row vector of the basis matrix of NMF represents a topic in the document set [9]. The number of topics extracted by NMF corresponds with that of the clusters to be determined by the above approach. While gene data is categorized into a cluster [7], a document is not necessarily classified into only one topic but can be labeled as multiple topics. The previous clustering algorithm is called hard clustering and the latter is soft clustering. A coefficient matrix of NMF is also suitable for soft clustering, and we have proposed a consensus matrix to determine the number of topics using soft clustering.

In the next section, we have described the details of consensus hard clustering using the coefficient matrix of NMF and a determination of the number of clusters. In section 3, we present our proposed approach. Our experimental results with the text documents are given in section 4. Finally, the conclusions of this study are outlined in section 5.

2 Consensus Hard Clustering with NMF

In this section, we have described the details of consensus hard clustering using the coefficient matrix of NMF and a determination of the number of clusters.

2.1 NMF Algorithm for a Document Set

NMF approximately factorizes a matrix of which all of the elements are non-negative values into two matrices with their elements being non-negative values. When NMF is applied to a document set, it has been reported that each basis represents a topic included in a document set. Now, a V × N term-document matrix D is defined as:

$$\mathrm{D} = \begin{bmatrix} \mathrm{d}_{11} & \cdots & \mathrm{d}_{1N} \\ \vdots & \ddots & \vdots \\ \mathrm{d}_{V1} & \cdots & \mathrm{d}_{VN} \end{bmatrix} \tag{1}$$

where V and N denote the number of terms in a document set and the number of documents, respectively. An element d_{ij} denotes a weight of the *i*th term in the *j*th document. The *j*th document is represented with a vector $\mathbf{d}_j$ named a document vector by a vector space model [10] as follows:

$$\mathbf{d}_j = [\mathrm{d}_{1j} \quad \cdots \quad \mathrm{d}_{Vj}]' \tag{2}$$

where $[\cdot]'$ denotes a transposition of a vector and matrix. NMF approximately factorizes the term- document matrix D into two matrices W and H such as:

$$\mathrm{D} \approx \mathrm{WH} \tag{3}$$

where W is a V × r matrix containing the basis vectors **w** as its columns and H is a r × N matrix containing the coefficient vectors **h** as its rows. In addition, each basis vector **w** is reported to represent a topic included in a document set. The constant *r* denotes the number of topics and it is arbitrary determined. *r* is set to less than the number of document N.

In this paper, we use the NMF algorithm that minimizes the Frobenius norm [11]. Given a matrix D, the optimal factors W and H are defined as the cost function f(W,H) and minimized, it is denoted as:

$$\mathrm{f(W,H)} = |\mathrm{D} - \mathrm{WH}|_{\mathrm{F}} \qquad \text{s.t. } \mathrm{W,H} \geq 0 \tag{4}$$

where $|\cdot|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix, and $\mathrm{W,H} \geq 0$ means that all elements of W and H are non-negative. In order to minimize f(W,H), the following updates are iterated until f(W,H) converges:

$$\overline{\mathrm{H}}_{\mathrm{ij}} = \frac{(\mathrm{D'W})_{\mathrm{ij}}}{(\mathrm{HW'W})_{\mathrm{ij}}} \mathrm{H}_{\mathrm{ij}} \tag{5}$$

$$\overline{\mathrm{W}}_{\mathrm{ij}} = \frac{(\mathrm{D'H})_{\mathrm{ij}}}{(\mathrm{WH'H})_{\mathrm{ij}}} \mathrm{W}_{\mathrm{ij}} \tag{6}$$

where $\overline{\mathrm{H}}$ and $\overline{\mathrm{W}}$ denote the updated factors.

2.2 Consensus Hard Clustering Using the Coefficient Matrix

Clustering using the NMF algorithm is performed based on the coefficient matrix H. The element of H, H_{ij}, denotes a degree of relativity between a document $\mathbf{d}_j$, and a topic $\mathbf{w}_i$. The topic label k_i of a document $\mathbf{d}_i$ is defined as:

$$\mathrm{k}_j = \arg\max_i \mathrm{H}_{ij} \tag{7}$$

where k_j is labeled an index of the maximum value among a row of H. Hard clustering means the label of a document is determined only by a topic.

Consensus clustering repeatedly performs classifications several times and integrates their results. In consensus clustering using NMF, the initializations of W and H are performed randomly at each repetition so that a document does not have the same label at all the repetitions. Therefore, evaluation of the clustering is performed with consensus matrices at each repetition. The element of the consensus matrix denotes whether two documents have the same label or not. The consensus matrix of the *t*th repetition $\mathrm{M}^{(t)}$ is established at each clustering time where t denotes the repetition count and its elements take only a binary value such as:

$$\mathrm{M}_{\mathrm{ij}}^{(\mathrm{t})} = \begin{cases} 1 & \mathbf{d}_m \text{ and } \mathbf{d}_n \text{ have a same label} \\ 0 & \text{otherwize} \end{cases}. \tag{8}$$

The results of consensus hard clustering are established by an average consensus matrix M that is the average of the consensus matrices for all of the repetitions. The average consensus matrix is expressed as:

$$\mathrm{M} = \frac{1}{\mathrm{T}} \sum_{t} \mathrm{M}^{(t)} \tag{9}$$

where T denotes the number of repetitions and the elements of M denote the relative degree between the documents.

2.3 Determination of the Number of Topics

It is important in any topic extraction method to determine the number of topics. The approach when using the consensus matrix is based on the assumption that if the results of the classified documents to topics at each repetition are robust, the number of topics is optimized. Therefore, the algorithm in [7] gradually changes the number of topics and searches the optimized one that provides the most robust clustering by random initialization at every repetition.

In order to measure the robustness of the clustering, a Cophenetic Correlation Coefficient (CPCC) is used. The element of the consensus matrix M_{ij} is regarded as the proximity between the ith document and jth document and the cophenetic matrix Q is established and its element Q_{ij} is defined as the proximity level where the ith and jth documents are grouped in the same cluster for the first time. In addition, the means of M and Q, i.e. μ_M, μ_{Q}, are defined as:

$$\mu_{\mathrm{M}} = \frac{2}{\mathrm{N}(\mathrm{N}-1)} \sum_{i=1}^{\mathrm{N}-1} \sum_{j=i+1}^{\mathrm{N}} \mathrm{M}_{ij} \tag{10}$$

$$\mu_{\mathrm{Q}} = \frac{2}{\mathrm{N}(\mathrm{N}-1)} \sum_{i=1}^{\mathrm{N}-1} \sum_{j=i+1}^{\mathrm{N}} \mathrm{Q}_{ij} \tag{11}$$

Using μ_{M}, μ_{Q}, CPCC is defined as:

$$\mathrm{CPCC} = \frac{\sum_{i=1}^{\mathrm{N}-1} \sum_{j=i+1}^{\mathrm{N}} (\mathrm{M}_{ij} - \mu_{\mathrm{M}})(\mathrm{Q}_{ij} - \mu_{\mathrm{Q}})}{\sqrt{\left(\sum_{i=1}^{\mathrm{N}-1} \sum_{j=i+1}^{\mathrm{N}} \mathrm{M}_{ij}^2 - \mu_{\mathrm{M}}^2\right)\left(\sum_{i=1}^{\mathrm{N}-1} \sum_{j=i+1}^{\mathrm{N}} \mathrm{Q}_{ij}^2 - \mu_{\mathrm{Q}}^2\right)}}. \tag{12}$$

The value of CPCC lies in the range of [-1, 1], and an index value close to 1 indicates a good robustness for the clustering. In addition, Q is derived with the Weighted Pair Group Method with Averaging (WPGMA).

3 Consensus Soft Clustering Based on NMF

In this section, our approach using consensus soft clustering with NMF is described.

3.1 Consensus Soft Clustering Using the Coefficient Matrix

Soft clustering is one of the clustering method approaches and the label of a classified document is represented as a probability of which a document belongs to a topic. It is also reported that soft clustering is applied to consensus clustering [12]. Soft clustering with NMF uses the value of the coefficient matrix H, directly. When $\mathrm{H}^{(t)}$ is the coefficient matrix at the *t*th repetition in consensus clustering, our proposed method defines the probability of which a document $\mathbf{d}_j$ is assigned to a topic C_i by normalization with respect to a row expressed as:

$$\mathrm{p}^{(t)}\left(\mathrm{C}_i|\mathbf{d}_j\right) = \frac{1}{\sum_{i=1}^{r} \mathrm{H}_{ij}^{(t)}} \mathrm{H}_{ij}^{(t)} \tag{13}$$

where $\mathrm{p}^{(t)}\left(\mathrm{C}_i|\mathbf{d}_j\right)$ denotes the probability of which the *j*th document is assigned to a *i*th topic at the *t*th repetition. The probabilistic consensus matrix $\mathrm{P}^{(t)}$ consists of the probability of an assignment and its element $\mathrm{P}_{ij}^{(t)}$ is defined as:

$$\mathrm{P}_{ij}^{(t)} = \sum_{k=1}^{\mathrm{r}} \mathrm{p}^{(t)}(\mathrm{C}_k|\mathbf{d}_i)\mathrm{p}^{(t)}\left(\mathrm{C}_k|\mathbf{d}_j\right) \tag{14}$$

$\mathrm{P}_{ij}^{(t)}$ denotes the joint probability that both classified labels of the *i*th document and *j*th document are the same. In addition, it assumes that the *i*th and *j*th documents are independently assigned to *k*th class. Finally, the results of consensus soft clustering depend on the average of $\mathrm{P}^{(t)}$s for all of the repetitions expressed as:

$$\mathrm{P} = \frac{1}{\mathrm{T}} \sum_{t} \mathrm{P}^{(t)} \tag{15}$$

where P denotes the emerging consensus probability matrix of which element is the expected probability that two documents are assigned to the same topic.

3.2 Determination of the Number of Topics

In our proposed method, in order to determine the number of topics, the CPCC described in equation (12) is also used and the number of topics gradually changes to explore an optimized one. The derivation of the CPCC for our proposed method substitute the consensus probabilistic matrix P_{ij} for M_{ij} in equation (12). The cophenetic matrix Q is also derived from P_{ij}.

The algorithm of our proposed method to determine the number of topics is as follows:

```
1. Set the maximum number of topics K to explore and the repetition time T of the
   consensus clustering.
2. Allow X as the term-document matrix.
3. Iterate until $k = K$
      a. Iterate until $t = T$
            i. Initialize W and H to random positive matrices.
           ii. Derive W and H using NMF with rank k.
          iii. Calculate the $t$th consensus probabilistic matrix $P^{(t)}$ with H.
      b. Calculate the average of $P^{(\mathrm{t})}$s as a consensus probabilistic matrix P.
      c. Evaluate P using CPCC.
4. Determine the number of topics depending on the maximum and the transition of
   the CPCC values.
```

4 Experiments and Discussions

In this section, we will describe details of our experiments with documents and discuss the results.

4.1 Experiments with Text Documents

In this paper, documents are used as experimental data to confirm the effectiveness of the proposed consensus clustering. The documents used in this experiment belong to the test collection entitled "cluto"[1] developed for evaluations of text clustering. The details of the documents used in this experiment are presented in Table 1.

Table 1. Details of experimental text data

Data Name	# of Documents	# of Words	Given #
tr11	414	6,429	9
tr13	313	5,804	8
tr23	204	5,832	6
tr31	927	10,128	7
tr41	878	7,454	10
tr45	690	8,261	10

The field names of Table 1, i.e., "Data Name", "# of Documents", "# of Words", and "Given #", refer to the name of a document set, the number of documents belonging to each document set, the number of words included in each document set, and the number of clusters given by the data constructor, respectively.

[1] http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download

Before a clustering process, each document set in Table 1 was represented as a term-document matrix, as shown in equation (1). Each document in a document set was represented as a document vector in the term-document matrix. Its element d_{ij} was established as the weight of the *i*th word in the *j*th document of the document set and is derived by the tf-idf method with a normalization. The weight ω_{ij} by the tf-idf method is defined as:

$$\omega_{ij} = \mathrm{tf}_{ij} \log \frac{\mathrm{N}+1}{\mathrm{df}_i} \tag{16}$$

where tf_{ij} and df_j denote the *i*th term frequency in the *j*th document and document frequency, respectively. In addition, the reason why the number of documents is set to N+1 in the inverse document frequency is because the value of ω_{ij} becomes 0 when the *i*th term appears in all documents. The value of d_{ij} in equation (1) was the normalized value of ω_{ij} by the *i*th row, as follows:

$$\mathrm{d}_{ij} = \frac{\omega_{ij}}{\sqrt{\sum_i^{\mathrm{V}} \omega_{ij}^2}}. \tag{17}$$

After construction of a term-document matrix, it was applied to propose consensus clustering as the number of topics k gradually changed from 2 to 20. The repetition time of clustering for each number of topics was set to 50 repetitions, and the termination condition of NMF in each repetition was defined as the iteration time of NMF reached 100 times or the value of the Frobenius norm for the difference between D and WH converged under 10E-5.

Finally, the consensus matrix for each number of topics was evaluated by the CPCC value and the optimized number of topics searched. Evaluations were performed from the viewpoint of the transition of the CPCC values, and the number of topics of which the CPCC values takes the highest and local maximum. The number k_{lm} that takes a local maximum value of a CPCC is defined as follows:

$$k_{\mathrm{lm}} = \{k | \mathrm{CPCC}_k > \mathrm{CPCC}_{k-1} \cap \mathrm{CPCC}_k > \mathrm{CPCC}_{k+1}\} \tag{18}$$

where CPCC_k denotes the CPCC values where the number of topics is set to k. In addition, it was compared with the number given by the data constructor.

We also performed the approach to use hard consensus clustering as a comparative method.

4.2 Experimental Results

The experimental results are presented in this section and Fig.1 presents the transition of the CPCC values for each number of topics. In Fig.1, “trxx” denotes the data name and “Pro” and “Com” refers to the proposed method and comparative method, respectively. In addition, Table 2 shows the top 5 ranked numbers of topics with respect to the CPCC value for each document set with the proposed method and comparative one. In Table.2, “order” means the rank and “Given #” means the same as that in Table 1. The number in italic bold and underscored shows that the number of topics is equal to and neighbors the given number, respectively. In addition, “neighbor” is defined here as the range from -1 to +1 of the given number. Table 3 shows the numbers of topics of which the points take the local maximum values of CPCC. In Table 3, the “Local Max. Topic #” denotes the numbers of topics and “Given #” means the same as the field presented in Table 2.

4.3 Discussion

We first focused on the number of topics of which the CPCC values are highest in Table 2. Compared with the number of clusters given by the data constructor, only the document set named “tr13” was successful in estimating the given number of clusters for our proposed method. The top 5 ranked numbers with respect to the CPCC values required large numbers of topics for almost all of the document sets. It is natural in a sense that the more complicated a model becomes meaning that the number of topics is incremented, the more suitable the model is for the data. In addition, the CPCC values are high at the point of which the topic number is also set to a large number,

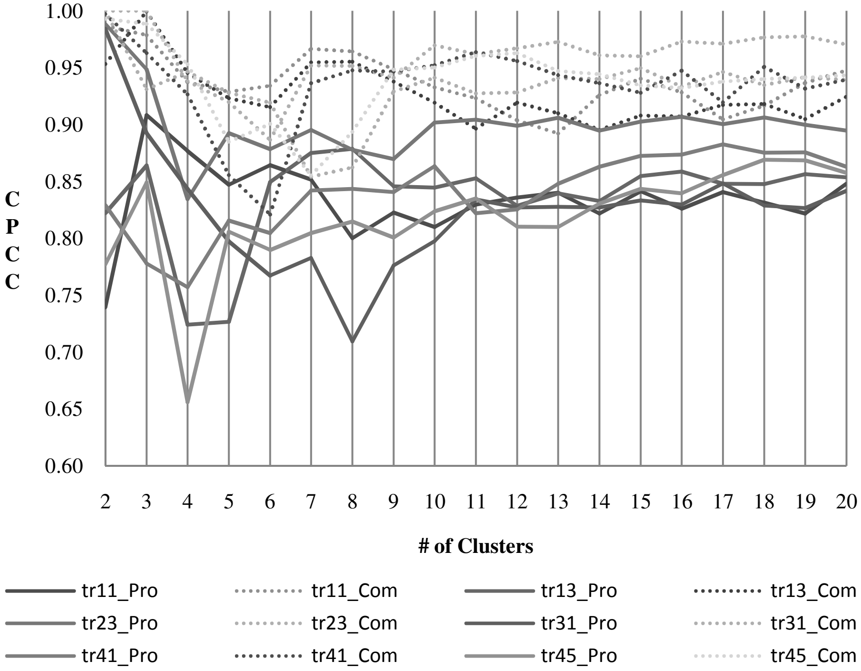


Fig. 1. Transition of CPCC values for each cluster size

Table 2. Top 5 ranked number of clusters by the CPCC value for each method

order	1	2	3	4	5	Given #
tr11_Pro	3	4	6	7	20	9
tr11_Com	2	3	7	8	***9***	
tr13_Pro	***8***	7	3	16	19	8
tr13_Com	3	***8***	7	2	4	
tr23_Pro	2	3	16	18	13	6
tr23_Com	2	3	19	18	16	
tr31_Pro	2	3	17	4	20	7
tr31_Com	2	15	4	17	19	
tr41_Pro	17	19	18	16	15	10
tr41_Com	2	11	3	12	***10***	
tr45_Pro	18	19	20	17	3	10
tr45_Com	2	3	12	11	4	

Table 3. Topic #s of CPCC's local maximum points

	Local Max. Topic #	Given #
tr11_Pro	3,6,***9***,13,15,17	9
tr11_Com	7,15	
tr13_Pro	3,***8***,11,13,16,19	8
tr13_Com	3,***8***,12,15,18	
tr23_Pro	5,7,11,13,16,18	6
tr23_Com	7,10,13,16,19	
tr31_Pro	***7***,11,13,15,17	7
tr31_Com	4,10,13,15,17,19	
tr41_Pro	5,8,***10***,17,19	10
tr41_Com	8,11,16,18	
tr45_Pro	3,5,8,11,15,18	10
tr45_Com	6,12	

however, when Fig. 1 is also considered, it is notable that the CPCC values hardly changes even though the number of topics is incremented over a certain number. The comparative method, i.e., using a hard clustering shows that the number of topics are equal to or neighbor a given number. However, the variance among the 5 ranked numbers is too large to actually select the best one. In addition, all of the top ranked numbers of the topics are 2 or 3, since the probability of assigning the same labels is considered to be high when the number of topics is small. For example, if the topic numbers are set to 2 and 3, the probability of which both $\mathbf{d}_m$ and $\mathbf{d}_n$ are assigned to a same topic is calculated as 1/2 and 1/3, respectively. In addition, if the range of the number of topics to explore is extended, the maximum value of the CPCC may be larger for a large number of topics. Therefore, it is difficult to determine the topic number depending on the maximum value of CPCC.

Focusing on Fig.1, the transition of the CPCC values by our proposed method requires a local maximum at a given number or its neighbors. Hence, we next focused on the number of topics of which the CPCC values are local maximum points in Table 3. Our proposed method shows that the 4 topic numbers of a 6 document set are equal to the given number and the other numbers neighbor the given number. Searching the local maximum points, our proposed method shows a topic number closer to the given number than that of a comparative method. Our proposed method is, thus, useful in refining the optimized topic numbers for actual usage. In contrast, the comparative method can hardly find the topic numbers equaling or neighboring the given number in this evaluation, though the given numbers were estimated by the maximum values of the CPCC.

5 Conclusions

We have proposed a novel approach to estimate the number of topics in a document set using consensus soft clustering based on NMF. We could confirm that our proposed method achieves refinement of the optimized number of topics in a document set for the actual document set and the difference in estimations between hard clustering and soft clustering.

Analysis of other document sets and approaches in consensus clustering algorithms are underway and will be discussed in future works.

References

1. Larsen, B., Aone, C.: Fast and Effective Text Mining using Linear-time Document Clustering. In: 5th International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 16–22 (1999)
2. Pelleg, D., Moore, A.: X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In: 17th International Conference on Machine Learning, pp. 727–734 (2000)
3. Windham, M., Culter, A.: Information Ratios for Validating Mixture Analysis. Journal of the American Statistical Association 87, 1182–1192 (1992)
4. The, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet Process. Technical Report 653, Department of Statistics, University of California at Berkeley (2004)
5. Monti, S., Tamayo, P., Mesirov, J., Golub, T.: Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Journal of Machine Learning 52, 91–118 (2003)
6. Li, T., Ding, C.: Weighted Consensus Clustering. In: Jonker, W., Petković, M. (eds.) SDM 2008. LNCS, vol. 5159, pp. 798–809. Springer, Heidelberg (2008)
7. Brunet, J.P., Tamayo, P., Golub, T.R., Mesirov, J.P.: Metagenes and Molecular Pattern Discovery using Matrix Factorization. PNAS 101(12), 4164–4169 (2004)
8. Rui, X., Wunsch II, D.C.: Clustering, pp. 267–268. J. Wiley & Sons Inc., NJ (2009)

9. Berry, M.W., Browne, M., Langville, A.N.: Algorithms and Applications for Approximate Nonnegative Matrix Factorization, V. In: Pauca, V.P., Plemmons, R.J. (eds.) Computational Statistics & Data Analysis, vol. 52(1), pp. 155–173 (2008)
10. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill Book Company, New York (1983)
11. Lee, D.D., Seung, H.S.: Algorithms for Non-negative Matrix Factorization. Advanced Neural Information Processing Systems 13, 556–562 (2001)
12. Punera, K., Ghosh, J.: Consensus-Based Ensembles of Soft Clustering. In: International Conference on Machine Learning: Models, Technologies & Applications (MLMTA 2007), pp. 3–9 (2007)

Infrastructure Aided Privacy Preserving-Authentication in VANETs

Brijesh Kumar Chaurasia[1], Shekhar Verma[1], and G.S. Tomar[2]

[1] Indian Institute of Information Technology, Allahabad, India
{bkchaurasia,sverma}@iiita.ac.in
[2] Malwa Institute of Technology and Management, Gwalior, India
gstomar@ieee.org

Abstract. The paper presents a privacy preserving authentication protocol for vehicles in a VANET. The authentication process involves authentication of the vehicle and the corresponding RSU by a fixed infrastructure (CTA). Every RSU has a public key infrastructure and continuously broadcasts public key. The vehicle encrypts its identity, RSU Id and timestamp using its symmetric key. The encrypted bits along with the pseudonym of vehicle and timestamp are again encrypted by the public key of RSU and send to the RSU. RSU forwards the encrypted part to the CTA. CTA sends its verification to the RSU. The verification of the vehicle is encrypted by the vehicle symmetric key and sends along with authentication of the vehicle to the RSU. The encrypted portion is forwarded to the vehicle which confirms the authentication of the RSU after decryption. The CTA also sends a temporary short certificate to the vehicle for vehicle to vehicle communication. The whole process needs only one request and reply between different entities. Simulation results indicate the time taken ($\sim 223\, ms$) for the whole process is small and constitutes only a small portion of the stay time of a vehicle within an RSU region.

Keywords: Mutual authentication, public-private key, VANETs, Vehicles, Road Side Units.

1 Introduction

VANETs is a network of vehicles are moving on the road exchanging information. The network membership is very volatile with members joining / leaving a neighborhood as they move on the road. Vehicles are equipped with an On Board Unit (OBU) that has an event data recorder (EDR), global positioning system (GPS), forward and backward radar, computing facility, and short range wireless interface [1]. A bandwidth of 75 MHz has been allocated in the 5.850-5.925 GHz band and vehicles use dedicated short range communications (DSRC) protocol for communication [2]. According to DSRC, each vehicle periodically broadcast information. DSRC classifies the five basic classes of applications; public safety application, traffic management, traveler information, freight/cargo transport, and transit. Messages class can be divided in two categories; safety and non-safety categories. The entails that (a vehicle) the source of message be authenticated beforehand joining the networks. Since

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 74–84, 2010.

message can cause / prevent life endangering situations, the variety of a message must be ascertained before an action is taken. In life threatening situations, the time for message authentication is almost negligible. Moreover, a malicious vehicle can broadcast with different identities. In vehicular network, DSRC [2], recommends the range of communication of vehicle is 500 meters and 1000 meters for road side infrastructure, a vehicle sends each message within 100-300 milliseconds time interval. However, if 50-200 vehicles present in the communication of road side infrastructure then network is high density network and in this case receiver vehicle will needs to verify near about 500-2000 messages per second. So main issue of authentication protocol is low communication overhead and fast verification.

At the time of authentication, identities of claimant vehicle must be hidden from a verifier vehicle and on the other hand, the authority should be able to trace the claimant vehicle or sender of a message by revealing its identity when required, such as liability investigation etc. So privacy must be preserve and privacy should be conditional.

Privacy preserving authentication can be achieved by using pseudonyms that are intimately linked to the original identity [1]. The pseudonym may be generated by the fixed infrastructure [3], [4] or by the vehicle itself [5]. They may be presorted [1] or downloaded from a trusted site periodically [6], [7]. During communication, the pseudonyms are switched periodically [8], or when required [9], in a crowd or a maximizing zone [10].For entity authentication public key infrastructure (PKI) [11], [12], [13] is deployed, where a large number of short-lived anonymous credentials is installed in the OBU. One of them is randomly selected used as the private key for digitally signing the messages. Verification is through the public key. However, detection of malicious sender is difficult. The CA has to exhaustively search a very large credential database to find the identity the compromised vehicle. Moreover, the security overhead is usually bigger than the useful message contents. Authentication can be done between two parties through exchange of certificates. This scheme [14] uses the short certificate based on temporary anonymous certified keys (TACKs) and uses group signature for tracing and revocation. A regional authority distributes certificates and certifies temporary key created by vehicles for authentication. Vehicles download CRLs certification revocation list to find for revoked entities [15]. Group based schemes [16], [17], [18], [19] provide anonymity as a receiver cannot distinguish a member from its group. Group based schemes achieve both privacy and authentication. However, group formation, maintenance, revocation need to be further studied [20]. To reduce the size of certificate revocation list and avoid the overheads of PKI, identity based with group signature scheme is proposed. The ID-based cryptosystem simplifies the certificate management process. The ID-based cryptosystem avoids certificate exchange and storage overheads of previous proposed schemes. However, their framework is limited by the strong dependence on the infrastructure for short lived pseudonym generation, which also renders the signaling overhead overwhelming. Timed efficient and Secure Vehicular Communication (TSVC) scheme [21] is also proposed for authentication. This scheme needs to perform symmetric MAC operation instead of any asymmetric operation at the verifier vehicle. Verification time is reduced but required tight time synchronization between vehicles. RAISE [22] is a RSU-aided scheme, responsible for verifying the authenticity of the messages sent from vehicles and for notifying the results back to vehicles. Where the message authentication code (MAC) can be used for inter vehicles authentication under the aid of a RSUs. The proposed scheme has less computation and communication overhead as compared to PKI-based and the group signature based schemes. However, this scheme is highly depend upon road side infrastructure, communication will be effected due network jammed because VANET is highly densely deployed and frequently disconnected network.

The rest of the paper is organized as follows. Section 2 describes the problem. In section 3, the architecture of VANETs is described. The protocol description is given in section 4. The scheme is evaluated through simulation and results are in section 5; section 6 concludes the work.

2 Problem Definition

A malicious vehicle can be an outsider or may be previously good vehicle. This malicious vehicle may inject false messages with different identities with dire consequences. This necessitates that messages and vehicles both are to be authenticated. Message authentication process needs to be repeated for each new message. Vehicle authentication should be done at the time of message sending to the verifier. At the time of communication, mutual authentication should be done for vehicle and RSU. This authentication must preserve the privacy of the vehicle. However, this privacy must be conditional. The true identity of a vehicle must be revealed if required by law. Finally, since the lifetime of a vehicle with an RSU is small, the authentication time should be almost negligible.

3 Architecture of VANETs

The architecture of a VANET is shown in Figure 1. It consists of national trusted authority $(\boldsymbol{TA})$, under this authority there are state level trusted authorities $(\boldsymbol{STA})$, city level trusted authorities $(\boldsymbol{CTA})$ are under in $\boldsymbol{STA}$. In every $\boldsymbol{CTA}$ there are many road side infrastructures $(\boldsymbol{RSUs})$ and there are vehicles moving on a road with an $\boldsymbol{RSU}$, they lateral motion is very restricted and the motion is unidirectional except at the junctions. A vehicle moving in a particular direction can move at different speeds and also pause. Vehicles can take velocity as variable or profile based etc & these vehicles may overtake one another. Since the transmission range of any vehicle is more than the total width of the road, this lateral motion has no effect on communications and can therefore be neglected. An $\boldsymbol{OBU}\ /\boldsymbol{RSU}$ is equipped with private key / public key which will provided by it's adjacent higher authority like $\boldsymbol{TA}$ distributes the keys and certificate to state level authorities. $\boldsymbol{STA}$ will play the role of key distributors to $\boldsymbol{CTA}$ and similarly $\boldsymbol{CTA}$ distributes the key and short certificates to road side infrastructures and vehicles. Each vehicle has equipped with storage area named as Tamper Proof Devices (TPD) to store different keys and for prevention and detection of tampering. A vehicle store many pseudonyms along with public / private key and short certificate which will be received at the time of authentication by $\boldsymbol{CTA}$ via the corresponding $\boldsymbol{RSU}$. When vehicle will come within the transmission range of a $\boldsymbol{RSU}$ it receives its public identity. All vehicles have used pseudonyms to preserve the privacy for vehicle during the broadcast. City level trusted authority $(\boldsymbol{CTA})$ plays the role as a key distributor. All vehicles register with $\boldsymbol{CTA}$ via any RSU or such as police station, government office, petrol pumps, and service stations etc. It is given one secret key for signing the messages, one shared secret key for communication between vehicles to city level trusted authority via $\boldsymbol{RSU}$. The vehicle will receive a short certificate by $\boldsymbol{CTA}$ via any $\boldsymbol{RSU}$ during authentication process. This short certificate can be used to authenticate claimant vehicle to verifier entity of network when infrastructure is absent.

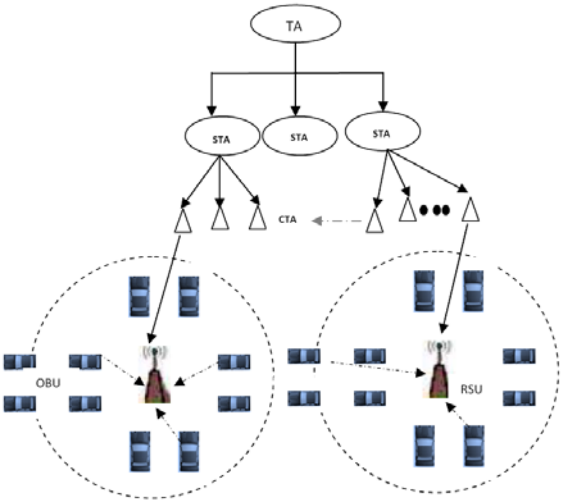


Fig. 1. Architecture of VANETs

Table 1. Notation used for this algorithm

Notation	Description
v_i	i^{th} Vehicle
TA	Trusted authority (National)
STA	Trusted authority (State)
RSU_i	i^{th} Road Side Infrastructure / Unit
CTA	Trusted authority (City)
PK_e^+	The Public key of any entity in vehicular network. Entity can be a vehicles or RSU_i etc.
$PK_{RSU_i}^+$	The Public key of i^{th} RSU_i
PK_v^+	The Private key of i^{th} v_i
$PK_{RSU_i}^-$	The Private key of i^{th} RSU_i
PK_v^-	The Private key of i^{th} v_i
PE_{RSU_i}	A public-key encryption function using the i^{th} RSU_i's public key
DE_{RSU_i}	A public-key decryption function using the i^{th} RSU_i's public key
K_{CTA}	The securely pre-shared symmetric key with CTA and vehicle
ID_v	Unique identity of vehicle, issued by CTA
ID_{RSU}	Unique identity of i^{th} road side infrastructure, issued by CTA
Sig_{CTA}	Signature of CTA

4 Protocol Description

The proposed authentication scheme is divided in four phases. The first three phases achieves the authentication with privacy using hashing and pseudonyms. Confidentiality is achieved by encryption of public identities and is in the fourth phase traceability and revocation is achieved by the simple cooperative scheme wherever infrastructure is absent or present. The proposed authentication scheme is as follows:

Phase I: Vehicle sends the request for association and authentication to RSU, phase II: The RSU forwards vehicle's authentication request to CTA, and Phase III: CTA sends the authenticated short certificate to vehicle via i^{th} RSU_i ; phase IV: Revocation.

Phase I: $v \rightarrow RSU_i$
Vehicle sends the request for association and authentication to RSU.
Step1. At the time of vehicle enters in the communication range of RSU, it receives RSU ID_{RSU} and $PK^{+}_{RSU_i}$ for sending authentication request.
Step2: Vehicle selects a pseudonym from its storage pool, and current time stamp t_0.
Step3: Computes a MAC value.

$$ps_0 = h(ID'_v, t_0)$$

t_0 is the 4 byte field time stamp for freshness to prevent message by replay attack / Sybil attack.
Step 4: Vehicle sends the authentication request to i^{th} RSU_i .
First, timestamp, vehicle identity and RSU identity are encrypted by the vehicle's shared key. Second, all values is again encrypted by public identity of RSU.

$$v_i \rightarrow RSU_i: PE_{ID_{RSU}}\{PE_{K_{CTA}}(ID_v , t_0, ID_{RSU_i}), ps_0, t_0 \}$$

Encryption technique is used to provide confidentiality.

Phase II: RSU forwards vehicle's authentication request to CTA.
Step 1: i^{th} RSU_i decrypt received association and authentication request and store ps_0, for the time duration until the CTA does not send the response to the RSU.
Step 2: i^{th} RSU_i will forward the encrypted packet to CTA .

$$RSU_i \rightarrow CTA : \{PE_{K_{CTA}}(ID_v , t_0, ID_{RSU_i}) \}$$

Phase III: CTA sends the authenticated short certificate to vehicle via i^{th} RSU_i .
Step1: CTA decrypts the authentication request by its shared key and verifies the vehicle and RSU_i.
Step2: After completion of the authentication process of vehicle and i^{th} RSU_i, CTA will issue the short certificate with time to live (t_1) time stamp to vehicle via i^{th} RSU_i.

$$CTA \rightarrow RSU_i: PE_{ID_{RSU}}[PE_{K_{CTA}}(ID_{RSU}, cert[Sig_{CTA}, t_1], ID_v)\, ps_0]$$

RSU_i will match the MAC value obtained from CTA from it's previously stored MAC valued if this is same then vehicle authenticates to RSU_i.
The certificate is valid for a time period determined by the approximate duration of stay of a typical vehicle in an RSU.

Step3: i^{th} RSU_i sends the authentication report to the vehicle. $RSU_i \rightarrow v$:

$$RSU_i \rightarrow v{:}\left[PE_{K_{CTA}}(ID_{RSU}, cert[Sig_{CTA}, t_1], ID_v)\ \right]$$

Vehicle receives the authentication certificate and at the same time vehicle will authenticate the RSU.

Phase IV: Revocation
Vehicle found some conditions regarding RSU that are such as:
(i) RSU_i is malicious, (ii) RSU_i is switched off, (iii) Large congestion in the network- RSU_i is overloaded, hence delay occurred, and (iv) CTA founds RSU_i is malicious at the time of authentication process.

(i) If any vehicle found the identity of i^{th} RSU_i was not valid then vehicle can inform the CTA connected by next RSU_{i+1} or connected by next other trusted infrastructure. So that CTA will verify the authenticity of that RSU_i. If finds RSU_i is malicious then broadcast the alert messages about RSU_i.
(ii) This condition is very rare in VANETs. If vehicle found i^{th} RSU_i is switched off then vehicle will report to next adjacent RSU_{i+1}. This will verify, if found true then broadcast the i^{th} RSU condition and to inform the CTA.
(iii) In this condition vehicle will be send association & authentication request again and again and wait some more time for authentication, otherwise resend association & authentication request to the next RSU_{i+1}. Vehicle will use the old certificate until didn't get new certificate.
(iv) If CTA founds that RSU_i is malicious then it will send information to vehicle and inform the other network entities or broadcast alert messages regarding false identity of RSU_i and also will take action to remove from the VANETs. CTA will listed this malicious RSU_i in revocation list, which stored in CTA. The revocation list can be seen at time to time by the connected from any type of infrastructure in VANETs such as next RSU_{i+1}, police station, government office , service stations and petrol pump etc. So this scheme is also able detect the false identity of network entities.

5 Simulation and Result Setup

5.1 Setup

In this section, simulation is conducted to verify the efficiency of the proposed secure protocol for inter vehicular communication applications with NCTUns [23]. For cryptographic delay we install MIRACL [24] and its library. So for these cryptographic delays we run a program that contains computational time of all standard hash function and encryption / decryption algorithms. The hardware/processor/clock of the system over which we install MIRACL is given in Figure 2.

Intel (R) Core (TM) @ Quad CPU
1.99 GB RAM
Q9300 @ 2.50 GHz

Fig. 2. CPU configuration over which we install MIRACL

We consider two types of different length of packets for authentication scheme. First when vehicle communicates to road side infrastructure then structure is as shown in figure 3a and when road side infrastructure responds to vehicle then structure is as shown in figure 3b. Lengths of packets are 108 bytes and 148 bytes respectably.

Type ID	Message ID	Payload	Time Stamp
2 byte	2 byte	100 byte	4 byte

Fig. 3a. Packet structure from RSU to OBU

Type ID	Message ID	Payload	Time Stamp	Signature
2 byte	2 byte	100 byte	4 byte	40 byte

Fig. 3b. Packet structure from OBU to RSU

In the simulation, speed of vehicle are (10-30 ms^{-1}), while communication range of VANET is 250-300 meters. Vehicle stays within transmission range of the *RSU* for a very short duration of time (approx. 10-20 sec.). In the proposed scheme there will be two types of delays one is communication delay and another is computational delay. For communication delay we have simulated in NCTUNs because in VANET environment communication protocol 802.11(p) is used. We have simulated number of vehicles 10, 20, 40 in fixed transmission range (300 m).

5.2 Setup

Data packets are generated at a constant bit rate at *RSU* as well as on OBU. Figure 4a and figure 4b, shows the average and maximum delay when number of vehicles varies from 5-40. Speed of vehicles are assumed here 10- 30 ms^{-1} and acceleration / deceleration = 3 ms^{-2}.

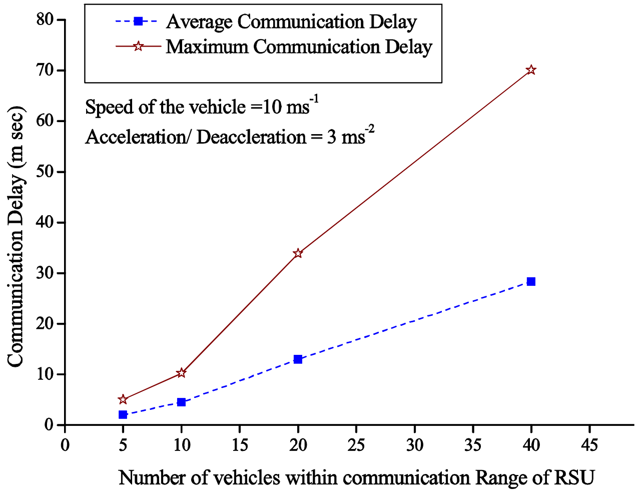


Fig. 4a. Average and Maximum communication delay at speed of source vehicle 10ms^{-1}

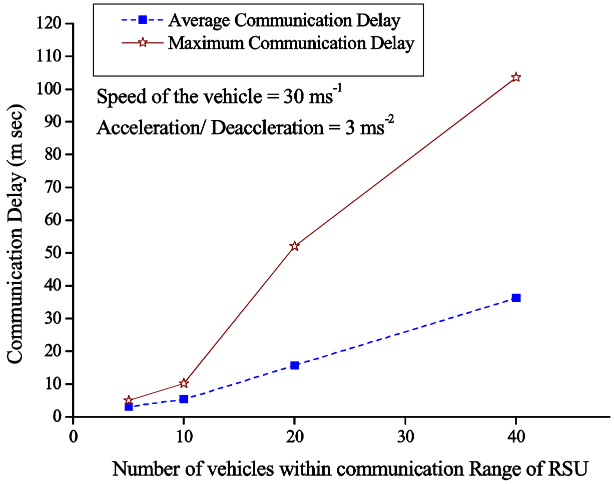


Fig. 4b. Average and Maximum communication delay at speed of source vehicle 30ms^{-1}

Computational Delay

For calculating delay of the authentication phase we analyze each step of the phase. Here we start with first step.

i. The delay when vehicle takes identity from RSU. t_2 is the time when vehicle send the request for authentication to RSU and RSU will send the public identity to vehicle $t_2 = t_0 + t_1$.

t_0 is the time when packet send by vehicle to RSU is 0.4081 ms.

t_1 is the communication delay of received packet from RSU when vehicles (around 40) are within the communication range of the RSU. Time t_1 is the maximum communication delay around 70ms and (~104) ms when vehicles having acceleration / deceleration of 3 ms^{-2} and speed of 10 ms^{-1} and 30 ms^{-1} respectively.

ii. t_3 is the delay when vehicle compute the hash function and encrypt the packet. Average delay of hash function (SHA-1) after multiple simulations is (~0.88) ms and similarly encryption or decryption delay t_4 is (~1.66) ms.

The delay of hash and encryption of the packet is $t_5 = t_3 + t_4$.

iii. Signing delay of the CTA is t_6 = (~1.33) ms. Verification delay t_7 is dependent on the computational delay of accessing the identity of claimant from its database and decryption delay t_4 .

Delay when RSU send the authentication request to CTA and CTA send the response to RSU along with the computational delay which is taken as 10 ms maximum.

Total time taken in authentication process is $T = t_2 + t_5 + t_7$.

Total maximum delay for authentication is *T* shown in figure 5. In figure 5a and figure 5b shown total maximum and average delay of the authentication process when number of vehicles varies 10 to 40 and speeds of vehicle is 10 ms^{-1} , and 30 ms^{-1} respectively with acceleration / deceleration taken as 3 ms^{-2}.

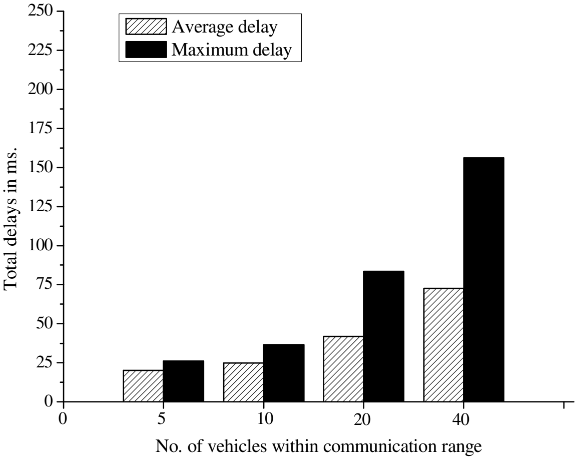


Fig. 5a. Average and maximum delay at speed of source vehicle 10ms^{-1}

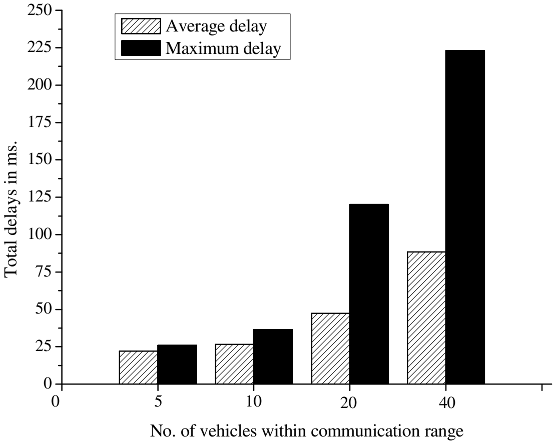


Fig. 5b. Average and maximum delay at speed of source vehicle 30ms^{-1}

6 Conclusion

In this paper, we provide a solution for privacy and mutual authentication process. We use the pseudonym based scheme for privacy preservation and mutual authentication.

The RSU was used as a mediator for authentication for both the RSU, itself, and requesting vehicle. Since the CTA is responsible for checking the credentials, the work of the RSU is drastically reduced. However, this requires all the RSU to be in continuous communication with a city level trusted authority, which may constitute a large overhead. This also solved the problem of malicious RSU along with the number of message exchange between different entities.

References

1. Dotzer, F.: Privacy Issues in Vehicular Ad Hoc Networks. In: Workshop on Privacy Enhancing Technologies, Dubrovnik, Cavtat, Croatia, pp. 197–209 (2005)
2. Dedicated Short Range Communications (DSRC), `http://www.leearmstrong.com/Dsrc/DSRCHomeset.htm`
3. Papadimitratos, P., Buttyan, L., Hubaux, J.-P., Kargl, F., Kung, A., Raya, M.: Architecture for Secure and Private Vehicular Communications. In: International Conference on ITS Telecommunications (ITST 2007), Sophia Antipolis, France, pp. 1–6 (2007)
4. Gerlach, M., Guttler, F.: Privacy in VANETs using Changing Pseudonyms - Ideal and Real (Poster Presentation). In: Proceedings of 65th Vehicular Technology Conference VTC 2007.Spring, Dublin, Ireland (2007)
5. Armknecht, F., Festag, A., Westhoff, D., Zang, K.: Cross-layer privacy enhancement and non-repudiation in vehicular communication. In: Proceedings of the 4th Workshop on Mobile Ad-Hoc Networks, WMAN 2007 (March 2007)
6. Raya, M., Hubaux, J.-P.: The Security of VANETs. In: VANET 2005, Cologne, Germany, pp. 93–94 (2005)
7. Ma, Z., Kargl, F., Weber, M.: Pseudonym-On-Demand: A New Pseudonym Refill Strategy for Vehicular Communications. In: Proc. IEEE 68th Vehicular Technology Conference, pp. 1–5 (2008)
8. Gruteser, M., Grunwald, D.: Enhancing location privacy in wireless LAN through disposable interface identifiers: a quantitative analysis. Paper presented in the Proceedings of ACM WMASH, pp. 46–55 (2003)
9. Chaurasia, B.-K., Verma, S., Tomar, G.-S., Abraham, A.: Optimizing Pseudonym Updation in Vehicular Ad-hoc Networks. In: Gavrilova, M.L., Tan, C.J.K., Moreno, E.D. (eds.) Transactions on Computational Science IV. LNCS, vol. 5430, pp. 136–148. Springer, Heidelberg (2009)
10. Chaurasia, B.-K., Verma, S.: Maximising Anonymity of a Vehicle. Inderscience, International Journal of Autonomous and Adaptive Communications Systems (IJAACS), Special Issue on: Security, Trust, and Privacy in DTN and Vehicular Communications 3(2), 198–216 (2010)
11. Raya, M., Hubaux, J.-P.: Securing Vehicular Ad Hoc Networks. Journal of Computer Security, Special Issue on Security, Ad Hoc and Sensor Networks 15(1), 39–68 (2007)
12. Hubaux, J.-P., Capkun, S., Luo, J.: The security and privacy of smart vehicles. IEEE Security & Privacy magazine 2(3), 49–55 (2004)
13. Raya, M., Hubaux, J.-P.: The security of vehicular ad hoc networks. In: Proceedings of the ACM Workshop on Security of Ad Hoc and Sensor Networks (SASN), pp. 11–21 (2005)
14. Studer, A., Shi, E., Bai, F., Perrig, A.: TACKing Together Efficient Authentication, Revocation, and Privacy in VANETs. In: IEEE SECON 2009, Rom, Italy, pp. 1–9 (2009)
15. Golle, P., Greene, D., Staddon, J.: Detecting and correcting malicious data in VANETs. In: Proceedings of VANET 2004, pp. 29–37 (2004)

16. Sampigethaya, K., Li, M., Huang, L., Poovendran, R.: AMOEBA: Robust Location Privacy Scheme for VANET. IEEE JSAC 25(8), 1569–1589 (2007)
17. Guo, J., Baugh, J.-P., Wang, S.: A Group Signature Based Secure and Privacy- Preserving Vehicular Communication Framework. In: Proc. of the Mobile Networking for Vehicular Environment (MOVE) workshop in conjunction with IEEE INFOCOM, pp. 103–108 (2007)
18. Calandriello, G., Papadimitratos, P., Lioy, A., Hubaux, J.-P.: Efficient and robust pseudonymous authentication in VANET. In: Proceedings of the Workshop on Vehicular Ad Hoc Networks (2007)
19. Lu, R., Lin, X., Zhu, H., Ho, P.-H., Shen, X.: ECPP: Efficient conditional privacy preservation protocol for secure vehicular communications. In: Proceedings of the Workshop INFOCOM (2008)
20. Verma, M., Huang, D.: SeGCom: Secure Group Communication in VANETs. In: 6th IEEE Consumer Communications and Networking Conference (CCNC 2009), pp. 1–5 (2009)
21. Lin, X., Sun, X., Wang, X., Zhang, C., Ho, P.-H., Shen, X.: TSVC: Timed Efficient and Secure Vehicular Communications with Privacy Preserving. IEEE Trans. on Wireless Communications 7(12), 4987–4998 (2009)
22. Zhang, C., Lin, X., Lu, R., Ho, P.-H.: RAISE: an efficient rsu-aided message authentication scheme in vehicular communication networks. In: Proc. IEEE ICC 2008, Beijing, China (2008)
23. http://nsl.csie.nctu.edu.tw/nctuns.html
24. Shamus Software Ltd. MIRACL, Multiprecision Integer and Rational Arithmetic C/C++ Library, http://indigo.ie/~mscott

Computational Analysis of a Power Distribution System with Petri Nets

Andrés Iglesias and Akemi Gálvez

Department of Applied Mathematics and Computational Sciences,
University of Cantabria, Avda. de los Castros,
s/n, E-39005, Santander, Spain
{iglesias,galveza}@unican.es
http://personales.unican.es/iglesias

Abstract. Petri nets provide a simple yet very intuitive graphical model for processes such as concurrency, parallelism and synchronization. Furthermore, they have a solid mathematical foundation and a number of analysis methods available. Due to these reasons, Petri nets are especially suited for the analysis of large and complex systems. In this paper we apply the Petri nets formalism to represent and analyze a power distribution system. To this aim, we report briefly a Mathematica package developed by the authors to deal with Petri nets. The package is subsequently applied to determine the possible scenarios of a failure of the system and some associated behaviors.

1 Introduction

The analysis of large and complex systems is still a challenge from the computational point of view. Many different approaches have been described during the last few years to tackle this issue. Among them, those based on Petri nets (PN) are gaining more and more popularity. Most of the PN interest lies on their ability to represent a number of events and states in a distributed, parallel, nondeterministic or stochastic system and to simulate accurately processes such as concurrency, sequentiality or asynchronous control [2, 3, 12]. In addition, they have a solid mathematical foundation and a number of analysis methods available: linear algebraic techniques to verify properties such as place invariants, transition invariants and reachability; graph analysis and state equations to analyze their dynamic behavior; simulation and Markov-chain analysis for performance evaluation, etc. As a consequence, Petri nets have been successfully applied to many interesting problems including finite-state machines, concurrent systems, multiprocessors, parallel and distributed computation, formal languages, communication protocols, software verification and validation and many others [4–6, 8, 11].

In this paper we apply the Petri nets formalism to analyze the behavior and all possible scenarios of a failure of a power distribution system. As it will be shown later on, the PN scheme allows us not only to characterize all situations leading to system failures but also to answer many related questions such as

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 85–98, 2010.

Table 1. Mathematical definition of a Petri net

A Petri net (PN) is a 5-tuple $\mathcal{PN} = (\mathbf{P}, \mathbf{T}, \mathbf{A}, \mathbf{W}, \mathcal{M}_0)$ comprised of:

- a finite set of places, $\mathbf{P} = \{P_1, \dots, P_m\}$,
- a finite set of transitions, $\mathbf{T} = \{t_1, \dots, t_n\}$,
- a set of arcs, $\mathbf{A} \subseteq (\mathbf{P} \times \mathbf{T}) \bigcup (\mathbf{T} \times \mathbf{P})$
- a weight function: $\mathbf{W} : \mathbf{A} \to \mathbb{N}^q \quad (q = \#(\mathbf{A}))$
- an initial marking: $\mathcal{M}_0 : \mathbf{P} \to \mathbb{N}^m$

If $\mathcal{PN}$ is a finite capacity net, it also contains:

- a set of capacities, $\mathbf{C} : \mathbf{P} \to \mathbb{N}^m$
- a finite collection of markings $\mathcal{M}_i : \mathbf{P} \to \mathbb{N}^m$

the number of steps required to achieve a specific system state or the actions involved in such a process. All these questions are properly addressed by using a *Mathematica* package briefly reported in this paper.

The structure of this paper is as follows: firstly, some basic concepts and definitions about Petri nets (mainly intended for those unfamiliar with this kind of methodology) are given in Section 2. Section 3 describes briefly a *Mathematica* package to deal with some specific classes of Petri nets. The package is subsequently applied in Section 4 to the representation and failure analysis of the power distribution system. Some conclusions and further remarks close the paper.

2 Basic Concepts and Definitions

A *Petri net* (PN) is a special kind of directed graph, together with an initial state called the initial marking (see Table 1 for the mathematical details). The graph of a PN is a bipartite graph containing *places* $\{P_1, \dots, P_m\}$ and *transitions* $\{t_1, \dots, t_n\}$. Figure 1 shows an example of a Petri net comprised of three places and six transitions. In graphical representation, places are usually displayed as circles while transitions appear as rectangular boxes. The graph also contains arcs either from a place P_i to a transition t_j (*input arcs* for t_j) or from a transition to a place (*output arcs* for t_j). These arcs are labeled with their weights (positive integers), with the meaning that an arc of weight w can be understood as a set of w parallel arcs of unity weight (whose labels are usually omitted). In Fig. 1 the input arcs from P_1 to t_3 and P_2 to t_4 and the output arc from t_1 to P_1 have weight 2, the rest having unity weight.

A *marking* (state) assigns to each place P_i a nonnegative integer, k_i. In this case, we say that P_i *is marked with* k_i *tokens*. Graphically, this idea is represented by k_i small black circles (tokens) in place P_i. In other words, places hold tokens to represent predicates about the world state or internal state. The presence or absence of a token in a place can indicate whether a condition associated with

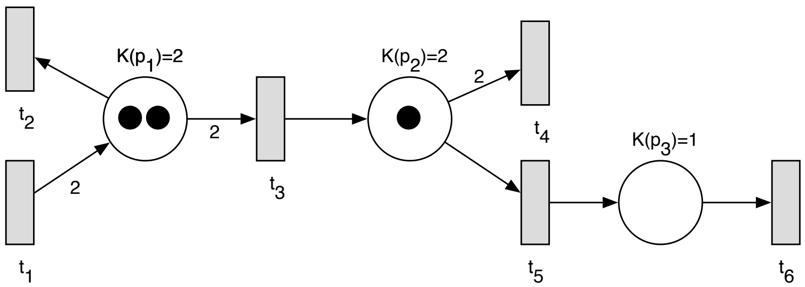


Fig. 1. Example of a Petri net

this place is true or false, for instance. For a place representing the availability of resources, the number of tokens in this place indicates the number of available resources. At any given time instance, the distribution of tokens on places, called Petri net marking, defines the current state of the modeled system. All markings are denoted by vectors $\mathcal{M}$ of length m (the total number of places in the net) such that the i-th component of $\mathcal{M}$ indicates the number of tokens in place P_i. From now on the initial marking will be denoted as $\mathcal{M}_0$. For instance, the initial marking (state) for the net in Figure 1 is $\{2, 1, 0\}$.

The pre- and post-sets of nodes are specified in this paper by a dot notation, where $\bullet u = \{v \in \mathbf{P} \bigcup \mathbf{T} / (v, u) \in \mathbf{A}\}$ is called the pre-set of u, and $u \bullet = \{v \in \mathbf{P} \bigcup \mathbf{T} / (u, v) \in \mathbf{A}\}$ is called the post-set of u. The pre-set of a place (transition) is the set of input transitions (places). The post-set of a place (transition) is the set of output transitions (places). The dynamical behavior of many systems can be expressed in terms of the system states of their Petri net. Such states are adequately described by the changes of markings of a PN according to a *firing rule* for the transitions: a transition t_j is said to be *enabled* in a marking $\mathcal{M}$ when all places in $\bullet t_j$ are marked. For instance, transitions t_2, t_3 and t_5 in Figure 1 are enabled, while transitions t_4 and t_6 are not. Note, for example, that transition t_4 has weight 2 while place P_2 has only 1 token, so arc from P_2 to t_4 is disabled. If transition t_j is enabled, it may or may not be fired (depending on whether or not the event represented by such a transition occurs). A firing of transition t_j removes $w_{i,j}$ tokens from each input place P_i of t_j and adds $w_{j,k}$ tokens to each output place P_k of t_j, $w_{j,k}$ being the weight of the arc from t_j to P_k. The fireability property of a transition t_j is denoted by $\mathcal{M}[t_j >$ while the creation of a new marking $\mathcal{M}'$ from $\mathcal{M}$ by firing t_j is denoted by $\mathcal{M}[t_j > \mathcal{M}'$.

A marking $\bar{\mathcal{M}}$ is *reachable* from any arbitrary marking $\mathcal{M}$ *iff* there exists a sequence of transitions $\sigma = t_1 t_2 t_3 \ldots t_n$ such that $\mathcal{M}[t_1 > \mathcal{M}_1[t_2 > \mathcal{M}_2 \ldots \mathcal{M}_{n-1}[t_n > \bar{\mathcal{M}}$. For short, we denote that the marking $\bar{\mathcal{M}}$ is reachable from $\mathcal{M}$ by $\mathcal{M}[\sigma > \bar{\mathcal{M}}$, where σ is called the *firing sequence*. The set of all markings reachable from $\mathcal{M}$ for a Petri net $\mathcal{PN}$ is denoted by $\biguplus[(\mathcal{PN}, \mathcal{M}) >$. Given a Petri net $\mathcal{PN}$, an initial marking $\mathcal{M}_0$ and any other marking $\mathcal{M}$, the problem of determining whether $\mathcal{M} \in \biguplus[(\mathcal{PN}, \mathcal{M}_0) >$ is known as the *reachability problem* for Petri nets. It has been shown that this problem is decidable [9] but it is also *EXP-time* and *EXP-space hard* in the general case [10]. In many practical applications it is interesting to know not only if a marking is reachable, but also what are the corresponding firing sequences leading to this marking. This can be done by using the so-called *reachability graph*, a graph consisting of the set of nodes of the original Petri net and a set of arcs connecting markings $\mathcal{M}_i$ and $\mathcal{M}_j$ *iff* $\exists t \in \mathbf{T}/\mathcal{M}_i[t > \mathcal{M}_j$.

A transition without any input place is called a *source transition*. Note that source transitions are always enabled. In Figure 1 there is only one source transition, namely t_1. A transition without any output place is called a *sink transition*. The reader will notice that the firing of a sink transition removes tokens but does not generate new tokens in the net. Sink transitions in Figure 1 are t_2, t_4 and t_6. A couple (P_i, t_j) is said to be a self-loop if $\mathbf{P}_i \in (\bullet t_j \bigcap t_j \bullet)$ (i.e., if P_i is both an input and an output place for transition t_j). A Petri net free of self-loops is called a *pure* net. In this paper, we will restrict exclusively to pure nets.

Some PN do not put any restriction on the number of tokens each place can hold. Such nets are usually referred to as *unfinite capacity net*. However, in most practical cases it is more reasonable to consider an upper limit to the number of tokens for a given place. That number is called the *capacity* of the place. If all places of a net have finite capacity, the net itself is referred to as a *finite capacity net*. All nets in this paper will belong to this later category. In such case, there is another condition to be fulfilled for any transition t_j to be enabled: the number of tokens at each output place of t_j must not exceed its capacity after firing t_j. For instance, transition t_1 in Figure 1 is initially disabled because place P_1 has already two tokens. If transitions t_2 and/or t_3 are applied more than once, the two tokens of place P_1 will be removed, so t_1 becomes enabled. Note also that transition t_3 cannot be fired initially more than once, as capacity of P_2 is 2.

3 A *Mathematica* Package for Petri Nets

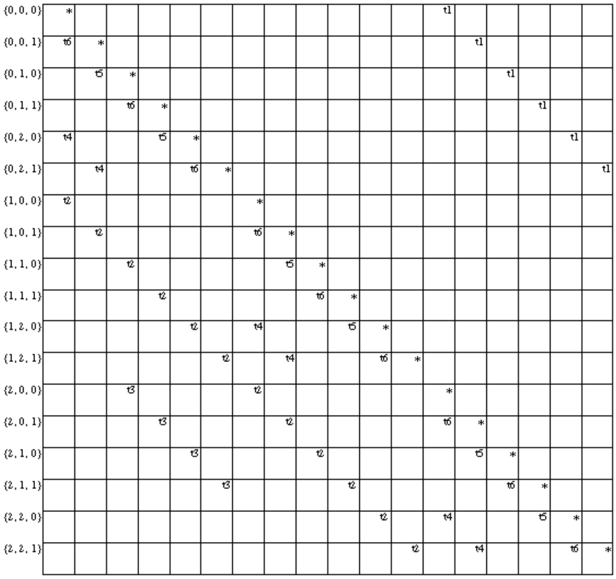
In this section a *Mathematica* package to deal with Petri nets is briefly reported (the reader is referred to [7] for a more detailed information about this software). We start our discussion by loading the package:

```
In[1]:= <<PetriNets‘
```

According to Table 1, a Petri net (like that in Figure 1 and denoted onwards as `net1`) is described as a collection of lists. In our representation, `net1` consists of three elements: a list of couples $\{place, capacity\}$, a list of transitions and a list of arcs from places to transitions along with its weights:

```
In[2]:= net1={{{p1,2},{p2,2},{p3,1}},{t1,t2,t3,t4,t5,t6},{{p1,t1,2},
    {p1,t2,-1},{p1,t3,-2},{p2,t3,1},{p2,t4,-2},{p2,t5,-1},{p3,t5,1},
    {p3,t6,-1}}};
```

Note that the arcs are represented by triplets $\{place, transition, weight\}$, where positive value for the weights mean output arcs and negative values denote input arcs. This notation is consistent with the fact that output arcs add tokens to the places while input arcs remove them. Now, given the initial marking $\{2, 1, 0\}$ and any transition, the `FireTransition` command returns the new marking obtained by firing such a transition:



{0,0,0}	*												t1					
{0,0,1}	t6	*												t1				
{0,1,0}		t5	*												t1			
{0,1,1}			t6	*												t1		
{0,2,0}	t4			t5	*												t1	
{0,2,1}		t4			t6	*												t1
{1,0,0}	t2						*											
{1,0,1}		t2					t6	*										
{1,1,0}			t2					t5	*									
{1,1,1}				t2					t6	*								
{1,2,0}					t2		t4			t5	*							
{1,2,1}						t2		t4			t6	*						
{2,0,0}			t3				t2						*					
{2,0,1}				t3				t2					t6	*				
{2,1,0}					t3				t2					t5	*			
{2,1,1}						t3				t2					t6	*		
{2,2,0}											t2		t4			t5	*	
{2,2,1}												t2		t4			t6	*

Fig. 2. Reachability graph for the Petri net in Figure 1

```
In[3]:= FireTransition[net1,{2,1,0},t2];
```

Out[3]:= {1,1,0}

Given a net and its initial marking, an interesting question is to determine whether or not a transition can be fired. The `EnabledTransitions` command returns the list of all enabled transitions for the given input:

```
In[4]:= EnabledTransitions[net1,{2,1,0}];
```

Out[4]:= {t2,t3,t5}

The `FireTransition` command allows us to compute the resulting markings obtained by applying these transitions onto the initial marking:

```
In[5]:=FireTransition[net1,{2,1,0},#]& /@ %;
```

Out[5]:= {{1,1,0},{0,2,0},{2,0,1}}

Note that, since transition $t1$ cannot be fired, an error message is returned:

```
In[6]:= FireTransition[net1,{2,1,0},t1];
```

Out[6]:= FireTransition: Disabled transition: t1 cannot be fired for the given net and the {2,1,0} marking.

From *Out[4]* and *Out[5]*, the reader can easily realize that successive applications of the `EnabledTransitions` and `FireTransition` commands allows us to obtain all possible markings and all possible firings at each marking. However, this is a tedious and time-consuming task to be done by hand. Usually, such markings and firings are graphically displayed in the reachability graph (see description above). The next input returns the reachability graph for our Petri net and its initial marking[1]:

[1] For an arbitrary PN the reachability graph may be of infinite size. This is not the case in this paper, as we restrict ourselves to finite PN.

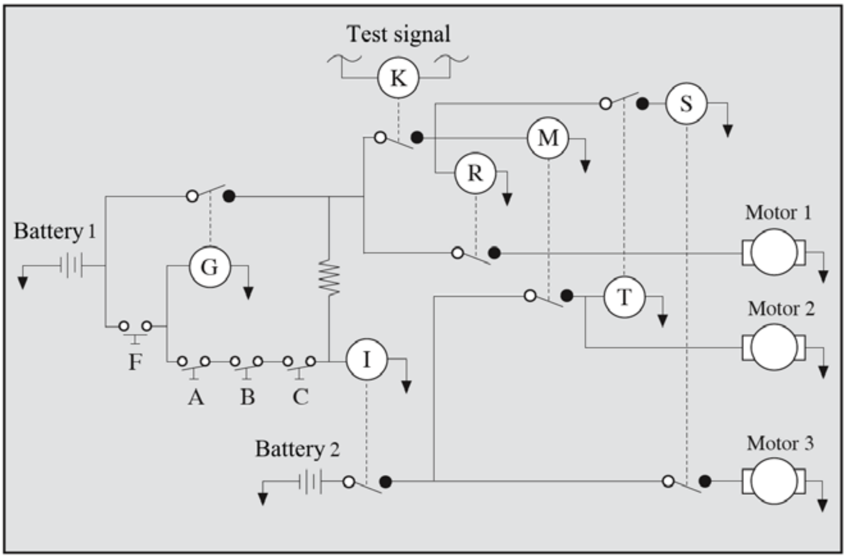


Fig. 3. Power distribution system diagram

```
In[7]:= ReachabilityGraph[net1,{2,1,0}];
```

Out[7]:= See Figure 2

Figure 2 can be interpreted as follows: the outer column on the left provides the list of all possible markings for the net. Their components are sorted in increasing order from the top to the bottom, according to the standard lexicographic order. For any marking, the row in front gives the collection of its enabled transitions. For instance, the enabled transitions for the initial marking $\{2, 1, 0\}$ are $\{t2, t3, t5\}$ (as expected from *Out[4]*), while they are $\{t1, t4, t6\}$ for $\{0, 2, 1\}$. Given a marking and one of its enabled transitions, we can determine the output marking of firing such transition by simply moving up/down in the transition column until reaching the star symbol: the marking in that row is the desired output. By this simple procedure, results such as *Out[5]* can readily be obtained.

4 Applying Petri Nets to a Power Distribution System

Petri nets have been successfully used as a formal method for the representation and analysis of large and complex dynamical systems during the last two decades. The reason is the large amount of mathematical tools available to analyse normal Petri nets [1, 12]. This allows us to perform a formal check of the properties related to the behavior of the underlying system, e.g., precedence relations amongst events, concurrent operations, appropriate synchronization, freedom from deadlock, repetitive activities, and mutual exclusion of shared

Table 2. Definition of variables

Q	EMF applied to Motor 2 for $t > 60$ seconds
N	M relay contacts remain closed for $t > 60$ seconds
J	I relay contacts fail to open when M contacts have been closed to $t > 60$ seconds
L	EMF remains on M coil for $t > 60$ seconds
H	G relay contacts fail to open when M contacts have been closed for $t > 60$ seconds
E	G relay contacts fail to open when K contacts have been closed for $t > 60$ seconds
D	EMF not removed from G relay coil when K contacts have been closed for $t > 60$ seconds
A, B, C	Primary failure of timer A, B, C (resp.)
F	Primary failure of pushbutton F
G, I, M	Primary failure of relays G, I, M (resp.)
K	Primary failure of test signal K

resources, to mention just a few. The ability of PN to verify the model formally is especially important for realtime safety-critical systems, such as air-traffic control systems, rail-traffic control systems, nuclear reactor control systems and so on. In this section, we apply Petri nets' formalism to failure detection of the system described in the next paragraphs.

4.1 Description of the System

Figure 3 shows a power distribution system with three motors, 1, 2 and 3, and three timers, A, B and C, which are normally closed. A momentary depression of pushbutton F applies power from a battery to the coils of cutthroat relays G and I. Thereupon G and I close and remain electrically latched. To check whether the three motors are operating properly, a 60-second test signal is impressed through K. Once K has been closed, power from battery 1 is applied to the coils of relays R and M. The closure of R starts motor 1. The closure of T applies power from battery 1 to coil S. The closure of S starts motor 3.

After an interval of 60 seconds, K is supposed to open, shutting down the operation of all three motors. Should K fail to be closed after the expiration of 60 seconds, all three timers A, B and C open, de-energizing the coil of G, thus shutting down the system. Suppose K open to de-energize G, and motor 1 stops. B and C act similarly to stop motor 2 or motor 3 should either M or S fail to be closed. For the sake of simplicity and brevity, in the following only the effect of motor 2 is analyzed.

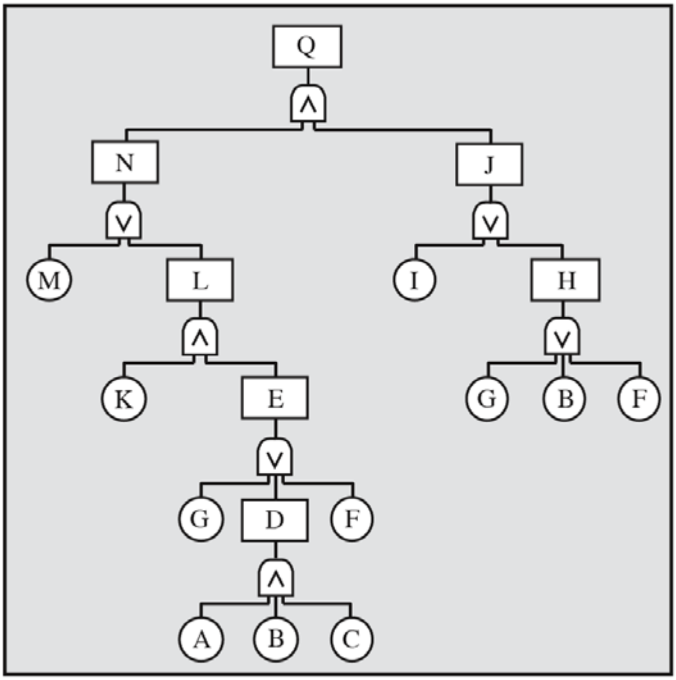


Fig. 4. Failure tree of the system

4.2 Failure Tree

To do so, let us consider Q to represent the event of failure of motor 2. We use the notation $Q = q$ to denote the failure of motor 2. Similarly, $a, b, c, \ldots$ represent the failures of the respective components $A, B, C, \ldots$. Based on the previous description of the system, we can derive the following logical expression for the failure of motor 2:

$$q = [m \vee (k \wedge g) \vee (k \wedge (a \wedge b \wedge c)) \vee (k \wedge f)] \wedge (i \vee g \vee b \vee f)$$

where the symbols $\vee$ and $\wedge$ denote the logical operators *or* and *and*, respectively. This expression is obtained by combining all possible partial failures of the components of the system that eventually produce a failure of motor 2. This equation can be expressed in a more intuitive way using the so-called failure tree. In this representation, the failures of timers A, B and C are combined into an intermediate cause of failure, D; then D is combined with G and F to define another intermediate cause of failure, E and so on. The resulting tree, depicted in Figure 4, includes the initial variables $\{A, B, C, F, G, I, K, M\}$ along with the intermediate failures $\{D, E, H, J, L, N\}$ that imply the failure of the motor. The final set of variables used in this example is $\{A, B, C, D, E, F, G, H, I, J, K, L, M, N, Q\}$. Their meaning is given in Table 2.

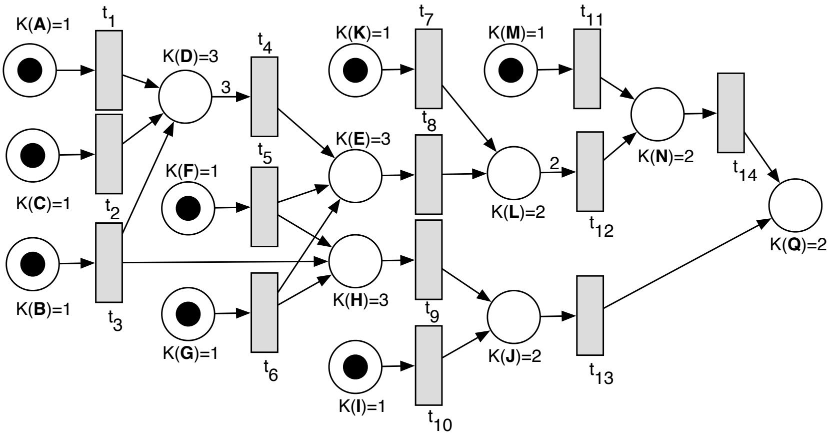


Fig. 5. Petri net of the power distribution system

4.3 Petri Net Representation

This system can be represented in *Mathematica* as:

```
In[8]:=motor={{{A,1},{C,1},{B,1},{D,3},{F,1},{G,1},{K,1},{E,3},{H,3},{I,1},
{M,1},{L,2},{J,2},{N,2},{Q,2}},{t1,t2,t3,t4,t5,t6,t7,t8,t9,t10,t11,t12,t13,
t14},{{t1,A,-1},{t1,D,1},{t2,C,-1},{t2,D,1},{t3,B,-1},{t3,D,1},{t3,H,1},
{t4,D,-3},{t4,E,1},{t5,F,-1},{t5,E,1},{t5,H,1},{t6,G,-1},{t6,E,1},{t6,H,1},
{t7,K,-1},{t7,L,1},{t8,E,-1},{t8,L,1},{t9,H,-1},{t9,J,1},{t10,I,-1},
{t10,J,1},{t11,M,-1},{t11,N,1},{t12,L,-2},{t12,N,1},{t13,J,-1},{t13,Q,1},
{t14,N,-1},{t14,Q,1}}};
```

leading to the Petri net displayed in Figure 5.

For illustrative purposes, the net is represented along with an initial marking represented by tokens in some places. In particular, the net in Figure 5 does represent the situation where all system components fail simultaneously (thus leading to a general failure of the system), given by the marking:

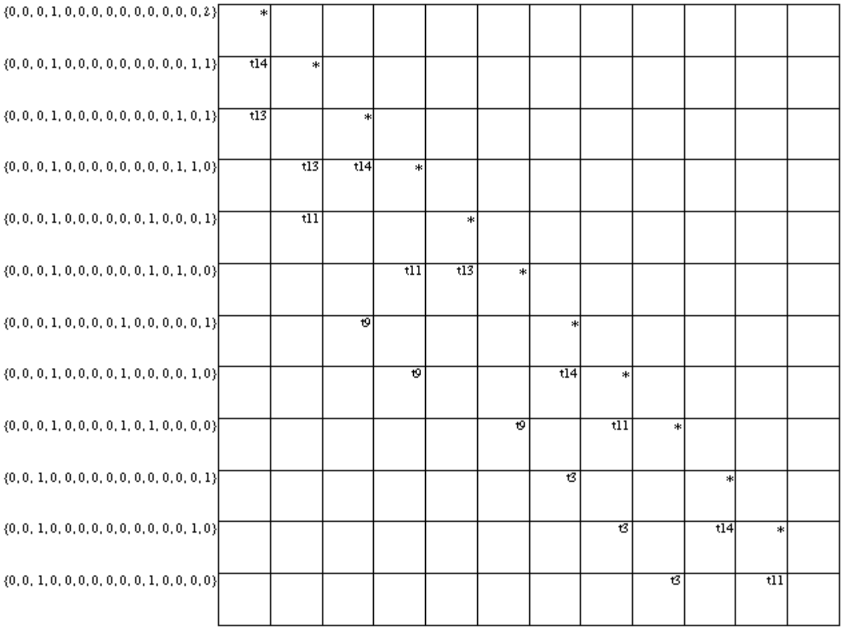
```
In[9]:=imk={1,1,1,0,1,1,1,0,0,1,1,0,0,0,0};
```

4.4 System Analysis

Identification of primary variables. From Fig. 4, it seems that the failure of the motor, Q, is a consequence of the failures of $\{A, B, C, D, E, F, G, H, I, J, K, L, M, N, Q\}$. However, some of these variables are intermediate variables, meaning that their failures are consequence of other primary variables (associated with the real components of the system). According to the description of the problem in terms of Petri nets, this means that Q is reached iff:

$$(t_{14} \vee t_{11} \vee t_{12} \vee (t_7 \wedge (t_8 \vee t_6 \vee t_5 \vee t_4 \vee (t_1 \wedge t_2 \wedge t_3)))) \wedge (t_{13} \vee t_{10} \vee t_9 \vee (t_6 \vee t_3 \vee t_5))$$

Note however that t_{14} is initially disabled because it requires either t_{11} or t_{12} to be fired. Note also that similar arguments can be applied to t_{13}, t_9 and t_8. At its turn, t_{12} is initially disabled because it requires both t_8 and t_7 to



{0,0,0,1,0,0,0,0,0,0,0,0,0,0,2}	*											
{0,0,0,1,0,0,0,0,0,0,0,0,0,1,1}	t14	*										
{0,0,0,1,0,0,0,0,0,0,0,0,1,0,1}	t13		*									
{0,0,0,1,0,0,0,0,0,0,0,0,1,1,0}		t13	t14	*								
{0,0,0,1,0,0,0,0,0,0,1,0,0,0,1}		t11			*							
{0,0,0,1,0,0,0,0,0,0,1,0,1,0,0}				t11	t13	*						
{0,0,0,1,0,0,0,0,1,0,0,0,0,0,1}			t9				*					
{0,0,0,1,0,0,0,0,1,0,0,0,0,1,0}				t9			t14	*				
{0,0,0,1,0,0,0,0,1,0,1,0,0,0,0}						t9		t11	*			
{0,0,1,0,0,0,0,0,0,0,0,0,0,0,1}							t3			*		
{0,0,1,0,0,0,0,0,0,0,0,0,0,1,0}								t3		t14	*	
{0,0,1,0,0,0,0,0,0,0,1,0,0,0,0}									t3		t11	

Fig. 6. Reachability graph of the power distribution system for failures of components B and M

be previously fired and the same applies to t_4. So, previous expression can be simplified to:
$(t_{11} \vee (t_7 \wedge (t_6 \vee t_5 \vee (t_1 \wedge t_2 \wedge t_3)))) \wedge (t_{10} \vee (t_6 \vee t_3 \vee t_5))$
that identifies the primary transitions of the system. This process can more easily be done by using the `EnabledTransitions` command:

```
In[10]:=EnabledTransitions[motor,imk]
```

Out[10]:= $\{t1,t2,t3,t5,t6,t7,t10,t11\}$

Reachable states of the system. This input returns the list of reachable markings (states) for the initial marking in Fig. 5:

```
In[11]:=lm=ListReachableMarkings[motor,%];
```

The output (stored in variable `lm`) is too large to be displayed here, as it includes 7074 possible states[2]:

```
In[12]:=Length[lm]
```

Out[12]:= 7074

However, the situation in which all system components fail at the same time is very rare. A more usual scenario is the failure of a specific component and then the discussion of the resulting cases simplifies greatly, as it reduces to a change

[2] Note that temporary states are also included in our analysis. Although they are associated with transient states of the system, they are often helpful in order to determine the sequences leading to system steady states.

in the initial marking, leading to different token locations (and hence, different states for the system) while the net topology no longer changes. In the following we will use several initial markings to account for failures of different system components. For instance, the failure of G and M yield 21 different states in the system:

```
In[13]:=fgm=ListReachableMarkings[motor,{0,0,0,0,0,1,0,0,0,0,1,0,0,0,0}];
Length[%]
Out[13]:= 21
```

of which only two lead to a failure of motor 2, represented by the condition $Q = 2$ (meaning that Q has two tokens):

```
In[14]:=Select[fgm,MatchQ[#,{___,2}]&]
Out[14]:= {{0,0,0,0,0,0,0,0,0,0,0,1,0,0,2},{0,0,0,0,0,0,0,1,0,0,0,0,0,0,2}}
```

while that the additional failure of B and M leads to the following reachability graph:

```
In[15]:=ReachabilityGraph[motor,{0,0,1,0,0,0,0,0,0,0,1,0,0,0,0}];
Out[15]:= See Figure 6
```

Analyzing transition sequences. One of the most interesting features of our approach is that it allows us to compute sequences of transitions leading to different states. For instance, we can determine the minimum number of transitions required for a given state to be reached. Because the Petri net is a planar graph, we implemented a modification of the minimal spanning tree algorithm. The command `MinimalTransitionSequence` admits three arguments: the Petri net, the initial marking and the target state and returns the list of transitions of minimal length yielding such target:

```
In[16]:=MinTransitionSequence[motor,imk,Q->2]
Out[16]:={t10,t11}
```

The obtained output means that there is only one possibility for the failure of motor 2 involving the minimum number of transitions. Note that t_{12} is not included in this solution, as it requires the prior firing of both t_7 and t_8, thus leading to larger sequences of transitions. Similarly, t_9 is not part of the solution to this problem because it requires the firing of either t_3, t_5 or t_6. This output means that the scenario of a system failure involving the minimum number of processes is that of a primary failure of relay M and the simultaneous occurrence of event J.

The target state in the previous command can be either a single condition (such as $Q = 2$) or a collection of multiple conditions involving Boolean operators. They are given in the form $lhs \rightarrow rhs$, according to the standard programming syntax in *Mathematica* for replacement rules. Thus:

```
In[17]:=MinimalTransitionSequence[motor,{0,0,1,0,0,0,0,0,0,0,1,0,0,0,0},
{I->0,Q->2}]
Out[17]:={{t3,t9,t11,t13,t14},{t3,t9,t11,t14,t13},{t3,t9,t13,t11,t14},
{t3,t11,t9,t13,t14},{t3,t11,t9,t14,t13},{t3,t11,t14,t9,t13},{t11,t3,t9,t13,t14},
{t11,t3,t9,t14,t13},{t11,t3,t14,t9,t13},{t11,t14,t3,t9,t13}}
```

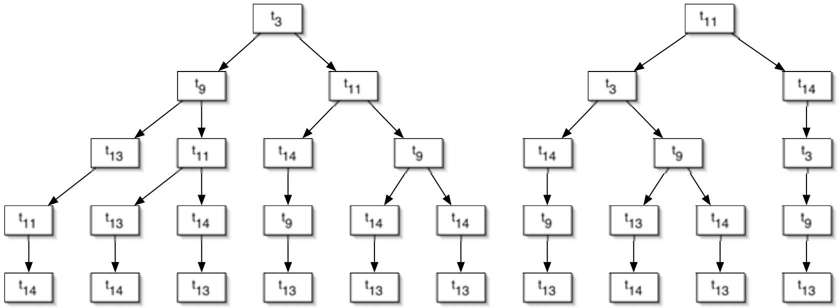


Fig. 7. Hierarchical tree of the sequences in *Out[17]*

returns the shortest sequences from the given marking that produces the failure of Q while the relay I is working properly. Such states are related to the failures of the following components:

```
In[18]:=PNComponents[motor,{0,0,1,0,0,0,0,0,0,0,1,0,0,0,0},#]& /@ %
```

Out[18]:=$\{\{H,I\},\{E,I\},\{G,I\},\{F,I\}\}$

Note that the initial marking already states that B and M fail simultaneously. The following input returns the same result graphically:

```
In[19]:=HierarchicalTree[%,Sort->Descending,ChildOrdering->LeftRight,
LayerOffset->Automatic,BoundingBox->Rectangle,TextAlign->{Middle,
Center}]
```

Out[19]:=See Figure 7

The information stored in the hierarchical tree of Figure 7 can also be expressed in terms of the processes executed during the runtime. The input:

```
In[20]:=DirectedGraph[%%,Field->Linear,BoundingBox->Oval,
Configuration->Polygonal,TextAlign->{Middle,Center},LocalStates->True]
```

Out[20]:=See Figure 8

returns the directed graph of the processes being executed according to the sequences in Figure 7. In this graph, failures of intermediate variables are denoted as **LF** (for local failures) and that the value $Q = 2$ yields the failure of the system (motor 2 in this example).

The interpretation of this directed graph is as follows: each node of the graph corresponds to a specific process denoted by a variable according to Table 2. The arrows indicate the flow of execution of such processes, as depicted in Figure 7. Note that some arrows are bidirectional, as some processes can be executed in reverse order as shown in Figure 7. By this simple procedure, all sequences described in the two hierarchical trees of Figure 7 can be embedded into the single graph of Figure 8. Note also that some other processes cannot be executed in reverse order, as they rely on previous steps. For instance, execution of transition t_9 does require the prior execution of transition t_3; this means that the processes associated with these transitions, namely, $H = 1$ and $J = 1$ respectively, must always be executed in that order and hence, reverse order is never allowed. This fact is represented by the directed arrow from $H = 1$ to $J = 1$. Similarly, the process $Q = 2$ (general failure of the system) always require the prior process

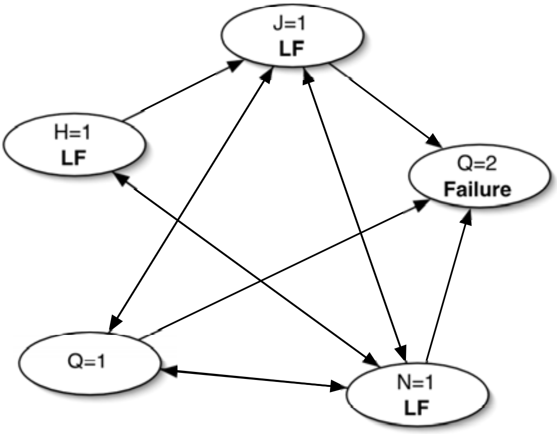


Fig. 8. Directed graph of the hierarchical tree of Figure 7

$Q = 1$ (failure of either N or J, the ultimate system components of Q) and thus, the arrow connecting both processes is undirectional. Therefore, this directed graph allows us to detect this kind of directional behavior at a glance. This feature is especially valuable for large systems, whose behavior is hard to be explained without the assistance of specialized tools.

5 Conclusions and Further Remarks

In this paper we have applied the Petri nets formalism to analyze the behavior of a power distribution system and obtain all possible situations of a failure of the system. To this aim, a *Mathematica* package developed by the authors has been successfully used. Our approach is very general in the sense that it can also be applied to the analysis of a bulk of larger and more complex examples not mentioned here because of limitations of space.

Regarding the implementation, it has been done in the computer algebra system *Mathematica* version 5.2. This software is available for all major operating systems (Windows, Macintosh, UNIX, Linux, etc.) on a variety of platforms and hardware configurations. All computations in this paper have been done on a Pentium IV, 3 GHz. with 1 GB. of RAM. Minimal requirements in terms of memory and storage capacity for our package do not exceed those commonly specified to install *Mathematica* thus facilitating the use of our package. Furthermore, all executions in our trials have been obtained in fractions of a second, making the package especially well suited for real-time applications and web services.

Future work includes the extension of our scheme to probabilistic systems and the inclusion of the system state equations in our analysis. In particular, the symbolic manipulation of the system state equations might lead to potential advantages for some problems such as the reachability, liveness and others.

This research has been supported by the Spanish Ministry of Education and Science, National Program in Computer Science (Project #TIN2006-13615) and the University of Cantabria.

References

1. Bourdeaud'huy, T., Hanafi, S., Yim, P.: Mathematical programming approach to the Petri nets reachability problem. European Journal of Operational Research 177, 176–197 (2007)
2. Gálvez, A., Iglesias, A., Corcuera, P.: Representation and Analysis of a Dynamical System with Petri Nets. In: Proc. International Conference on Convergence Information Technology-ICCIT 2007, Gyeongju, Korea, pp. 2009–2015. IEEE Computer Society Press, Los Alamitos (2007)
3. German, R.: Performance Analysis of Communication Systems with Non-Markovian Stochastic Petri Nets. John Wiley and Sons, Inc., New York (2000)
4. Iglesias, A.: A New Framework for Intelligent Semantic Web Services Based on GAIVAs. International Journal of Information Technology and Web Engineering, IJITWE 3(4), 30–58 (2008)
5. Iglesias, A.: Software Verification and Validation of Graphical Web Services in Digital 3D Worlds. In: Communications in Computer and Information Science. CCIS, vol. 56, pp. 293–300 (2009)
6. Iglesias, A.: Pure Petri Nets for Software Verification and Validation of Semantic Web Services in Graphical Worlds. International Journal of Future Generation Communication and Networking 3(1), 33–46 (2010)
7. Iglesias, A., Kapcak, S.: Symbolic computation of Petri nets. In: Shi, Y., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2007. LNCS, vol. 4488, pp. 235–242. Springer, Heidelberg (2007)
8. Iglesias, A., Luengo, F.: New Goal Selection Scheme for Behavioral Animation of Intelligent Virtual Agents. IEICE Transactions on Information and Systems, Special Issue on CyberWorlds E88-D(5), 865–871 (2005)
9. Kosaraju, S.R.: Decidability of reachability in vector addition systems. In: Proc. 14th Annual ACM Symp. Theory Computing, pp. 267–281 (1982)
10. Lipton, R.: The reachability problem requires exponential space.Technical report, Computer Science Department, Yale University (1976)
11. Luengo, F., Iglesias, A.: Designing an Action Selection Engine for Behavioral Animation of Intelligent Virtual Agents. In: Gervasi, O., Gavrilova, M.L., Kumar, V., Laganá, A., Lee, H.P., Mun, Y., Taniar, D., Tan, C.J.K. (eds.) ICCSA 2005. LNCS, vol. 3482, pp. 1157–1166. Springer, Heidelberg (2005)
12. Murata, T.: Petri nets: Properties, analysis and applications. Proceedings of the IEEE 77(4), 541–580 (1989)

Analysis of Effect of Different Factors on Burst Dropping Probability in Optical Burst Switching Network

Laxman D. Netak, Sandeep S. Udmale,
Girish V. Chowdhary, and Rupali Tajanpure

Department of Computer Engineering,
Dr. Babasaheb Ambedkar Technological University, Lonere, Raigad, M.S., India
ldnetak@yahoo.com,
sandeep_udmale1@yahoo.co.in,
girish.chowdhary@gmail.com,
rupalidixit18@gmail.com

Abstract. With an explosive growth of the Internet, optical network seems to be the candidate solution for future high-speed backbone networks. Optical burst switching (OBS) has been proposed as promising switching technology for the next generation wavelength division multiplexing (WDM) based network. The main issue in OBS based network is to reduce burst dropping probability. We classify different factors that affect burst dropping probability based on the schemes in which they are used. Schemes are like scheduling algorithm, burst assembly and burst reservation and factors are void filling and non-void filling scheduling algorithm, burst size, burst assembly schemes, offset time and different classes of traffic. This paper studies an impact of all above mention factors on to burst dropping probability with the help of extensive simulation on NS2. Simulation results shows that, different parameter must be properly used in the network.

Keywords: Wavelength division multiplexing (WDM), Optical burst switching (OBS), Scheduling.

1 Introduction

WDM technology has the enormous amount of bandwidth available in fiber cable. In WDM system, each fiber carries multiple communication channels and each channel operating on different wavelength. Such an optical transmission system has a potential capacity to provide tera bytes of bandwidth on a single fiber. WDM technology has the capability to provide the bandwidth for the increase in the huge on traffic demand of various application like audio, video and multimedia, which needs the QoS over the network [1].

The currently existing switching techniques can be broadly classified into optical circuit switching (OCS), optical packet switching (OPS) and OBS techniques [1], [2]. In OCS, an end-to-end optical lightpath is setup using a dedicated wavelength on each link from source to destination to avoid optical to electronic

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 99–113, 2010.

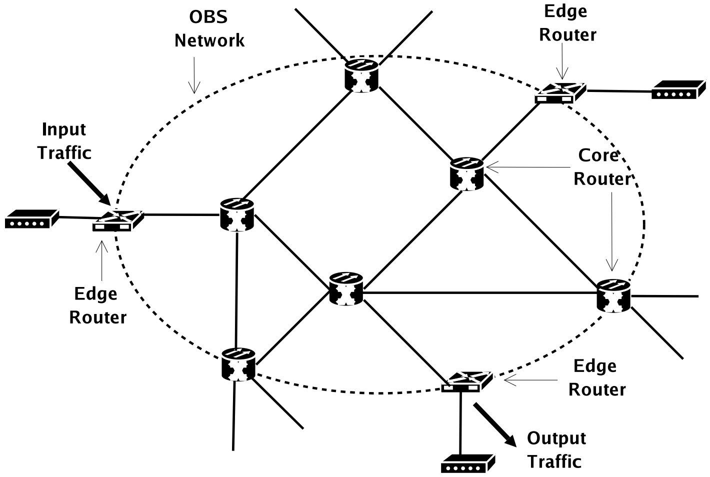


Fig. 1. OBS network model

(O/E/O) conversion at each intermediate nodes. Once the lightpath is setup, data remain in optical domain throughout transmission of data. OCS is relatively easy to implement but main drawback of OCS is circuit setup time and improper holding time of resources like bandwidth. On other hand, no circuit setup is required in OPS but packet header need to be processed in the electronic domain on hop-by-hop basis. Due to which data payload must wait in optical buffers like fiber delay lines (FDLs), which is very complex and challenging task in high speed optical networks. To do this task, OPS requires optical buffers, O/E/O converters and synchronizers. The new switching technology, which combines the merits of coarse gained OCS and fined gained OPS was proposed and called as OBS [1], [2], [3], [4].

In OBS network model, as shown in the Fig. 1 [5], there are two type of routers, edge and core router, which are connected by WDM links. Various type of client‘s data with same destination are aggregated at the edge router in a data burst. The data could be IP/SONET/SDH/ATM cell or combination of all packet type. In OBS, edge router is responsible for burst assembly/disassembly, scheduling of burst, transmission of burst, deciding the offset time, generation of burst control packet (CP) functions. Core router will forward the burst to its destination node [6], [7]. In OBS, a burst consist of header and payload called data burst. A burst header is called as CP. Typically, CP contains information about burst size and burst arrival time. The CP and payload are send separately on different channels called as control and data channel respectively as shown in Fig. 2 [8]. The burst is preceded in time by a CP, which is send on separate control wavelength. The preceded time is called as “offset time”.

After a burst is generated, the burst is buffered in the queue at edge router for an offset time before being transmitted to give its CP enough time to reserve network resource along its route. During offset time, packets belonging

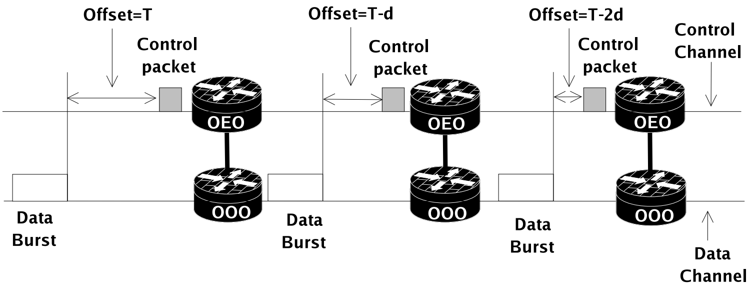


Fig. 2. Separate transmission of data and control signals

to that queue may continue to arrive. These extra packets are dropped [9]. At each intermediate node, CP undergoes O/E/O conversion to get it processed electronically. The time taken for processing a CP is called as the "processing time" [9], [10], [11]. Depending upon CP information wavelength is reserved for the incoming burst for that duration by core router [4], [6], [7], [8].

Basically, there are three different assembly schemes, namely threshold-based, timer-based and hybrid-based [9], [10].

In a timer-based scheme, a timer is started to initialize the assembly process. A burst containing all the packets in the buffer is generated when the timer exceeds the burst assembly period [9]. While in a threshold-based scheme, a burst is created and send into the OBS network when the total size of the packets in the queue reaches threshold value [9].

Hybrid assembly scheme is the combination of both threshold-based and timer-based assembly scheme [9]. In this scheme, a burst can be send out when either the burst length exceeds the desirable threshold value or the timer expire.

In OBS network, different wavelength reservation schemes are used for reserving the wavelength. One is called as Tell-And-Wait (TAW). In TAW, when source has the burst to send, it first reserve the wavelength along the route by sending "request" message. If the wavelength is granted by intermediate nodes along its route, a positive acknowledgment (PACK) message returns to source from the destination; otherwise negative acknowledgment (NACK) is received at source [4], [12], [13], [14].

Second scheme is called Tell-And-Go (TAG), in which two reservation schemes has been proposed. They are Just-Enough-Time (JET) and Just-In-Time (JIT). In JET, reservation is made by using CP information. The reservation is made for the duration of data burst. The resources are reserved and released implicitly. In JIT, the resources are reserved as soon as CP is received and hold resources until burst departure time. The resources are released explicitly by sending another control message and which results in bad resource utilization. Due to this the wavelength holding time to that node is larger than burst transmission time [4], [12], [13], [15].

2 Factors Affecting Burst Dropping Probability

In this section, we describe the different factors that affect burst dropping probability in OBS.

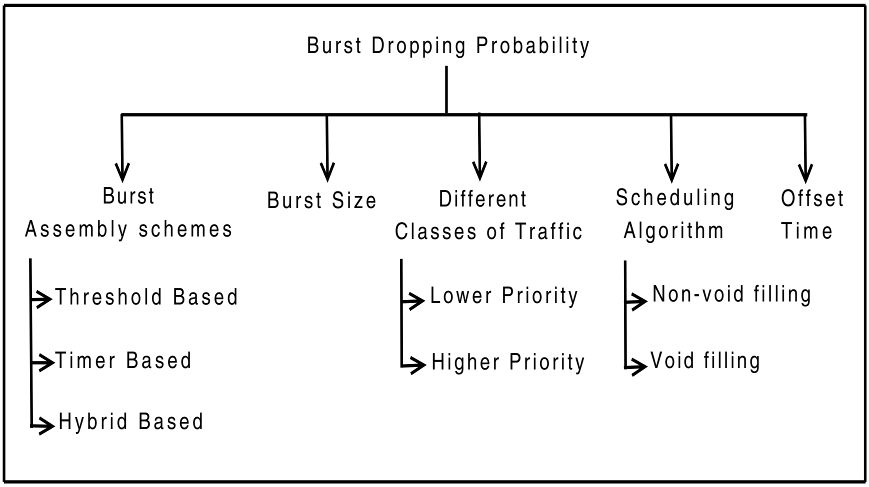


Fig. 3. Classification of various burst dropping factors

2.1 Burst Size and Burst Assembly Schemes

An assembly of the burst at the edge router is a challenging issue in the OBS, as it affect the network traffic.

To choose the appropriate time-out or threshold value for creating a burst is still an open issue. A smaller assembly granularity leads to higher number of burst and higher number of contentions, but the average number of packets lost per contention is less and it also increases the number of CPs. If the reconfiguration period of optical switching fabric is not negligible, smaller assembly granularity will lead to lower network utilization because each switching burst need reconfiguration period. On the other hand, a higher assembly granularity will lead to higher burst assembly delay and the average number of packets lost per contention is larger. There is a tradeoff between the number of contentions and the average number of packets lost per contention. The selection of optimal assembly granularity is strongly correlated to the type of input packet traffic. Due to great variability in the bursts size, OBS network can be viewed as a sandwiched between OPS and OCS network. That is, when burst size is small then it's equal to duration of an optical packet, OBS network can be seen as resembling an OPS network. On other hand, when burst size is large then OBS network can be seen as resembling an OCS network [9].

The limitation of the threshold-based scheme is that it does not provide any guarantee on the assembly delay that packet will experience. The limitation of the timer-based scheme is that it does not provide any guarantee on the size of burst.To overcome the drawback of timer-based and threshold based scheme, hybrid scheme has been proposed.

2.2 Types of Classes and Offset Time

Various traffic classes exists in OBS. To provide the QoS mechanism in OBS, higher priority classes are suggested to have the higher offset time value. Due to which resources will be reserved well before the actual burst arrive. Offset time value decreases with decrease in priority classes [13].

The value of the offset time need to be decided in such way that CP reach the destination well before arrival of burst. If the offset value is enough, the dropping probability of burst get reduce, but burst waiting time at the edge node increases. If the offset value is not enough, the dropping probability of burst get increases, but burst waiting time at the edge node decreases.

Hence, how to choose the appropriate value for offset time is still an open issue in OBS.

2.3 Burst Scheduling Algrithms

Another important factor which affects the network traffic is scheduling algorithms used to schedule burst. Arrival of bursts at OBS node is dynamic. Scheduling technique must schedule arrival burst on the available wavelengths for the entire duration of burst transmission. Scheduling technique must schedule burst efficiently and quickly. Scheduling algorithm should be able to process the CP fast enough before the burst arrives to the node. It should also be able to find proper void for an incoming burst to increase channel bandwidth utilization. Following are proposed burst scheduling algorithms in the literature:

Latest Available Unused Channel (LAUC) Algorithm [16], [17]. In LAUC, burst scheduling is done by selecting the latest available unscheduled data channel for each arriving data burst. In this algorithm, a scheduler keeps track of horizon for each channel. Horizon is the time after which no reservation has been made on that channel. LAUC searches the wavelength by using horizon information on each channel. The scheduler assign each arriving new burst to the data channel with minimum void formed by that burst on data channel.

As shown in Fig. 4, channel0 and channel1 is unscheduled channel at the arrival time *t* of the new burst. Channel0 will be selected for the new burst because the void generated on channel0 will be smaller than the void that would have been created if channel1 is selected.

Latest Available Unused Channel with Void Filling (LAUC-VF) Algorithm [16], [17]. In LAUC, the voids are created between two data burst assignment on the same data channel. This is termed as unused channel capacity. LAUC-VF is variant of LAUC. In this algorithm, a scheduler keeps track of horizon and voids for each channel. LAUC-VF maintain start and end time of void for each data channel. LAUC-VF searches for the void such way that newly formed void is very small compared to other voids.

As shown in Fig. 4, channel2, channel3 and channel4 is unscheduled channel at the arrival time t of the new burst. Channel2 will be selected for the new burst because void that will be produced between the bursts and coming data burst is the minimum void as compared to the channel3 and channel4.

Best-Fit (BF) Algorithm [16]. In BF, a scheduler keeps track of horizon and void for each channel. It also maintain start time and end time of void for each data channel. Scheduler tries to search for a void such way that newly created void is the smallest void before and after scheduled burst.

As shown in Fig. 4, channel2, channel3 and channel4 is unscheduled channel at the arrival time t of the new burst. Channel4 will be selected for the new burst because starting and ending void that will be produced between the bursts and coming data burst is the minimum void as compared to the channel2 and channel3.

Minimum Starting Void (Min-SV) Algorithm [16], [17]. In Min-SV, a scheduler keeps track of horizon and void for each channel. It also maintain start and end time of void for each data channel. Scheduler tries to search for a void such way that newly created void is the smallest void after scheduled burst.

As shown in Fig. 4, channel2, channel3 and channel4 is unscheduled channel at the arrival time t of the new burst. Channel2 will be selected for the new burst because starting void that will be produced between the bursts and coming data burst is the minimum void as compared to the channel3 and channel4.

Minimum Ending Void (Min-EV) Algorithm [16], [17]. In Min-EV, a scheduler keeps track of horizon and void for each channel. It also maintain start and end time of void for each data channel. Scheduler tries to search for a void such that newly created void is the smallest void before scheduled burst.

As shown in Fig. 4, channel2, channel3 and channel4 is unscheduled channel at the arrival time t of the new burst. Channel3 will be selected for the new burst because ending void that will be produced between the bursts and coming data burst is the minimum void as compared to the channel2 and channel4.

BF, Min-SV and Min-EV algorithms are the variant of LAUC-VF algorithm. All the void filling scheduling algorithm yields better bandwidth utilization and burst loss rate than LAUC algorithm. But all the void filling scheduling algorithm has a longer execution time than LAUC algorithm.

Fig. 4 shows the comparison of different scheduling algorithm [16] and Table 1 summarizes the above discussion using the following notations

- W : Number of wavelengths at each output port.
- m : Maximum number of data bursts(or reservations) on all channels.

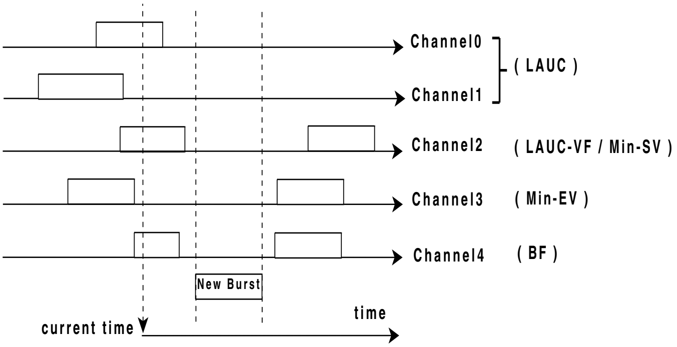


Fig. 4. An example showing how a new burst is scheduled by using different scheduling algorithm

Table 1. Comparison of different scheduling algorithm

Scheduling Algorithms	**Time Complexity**	**State Information**	**Bandwidth Utilization**
LAUC	O(W)	$Horizon_i$	Low
LAUC-VF	O(Wlogm)	$S_{i,j}$, $E_{i,j}$	High
BF	O(Wlogm)	$S_{i,j}$, $E_{i,j}$	High
Min-SV	O(logm)	$S_{i,j}$, $E_{i,j}$	High
Min-EV	O(logm)	$S_{i,j}$, $E_{i,j}$	High

- $Horizon_i$: Horizon of the i^{th} data channel.
- $S_{i,j}$ and $E_{i,j}$: Starting and ending time of j^{th} reservation on channel i.

3 Experiment Results

In order to evaluate the performance of OBS with burst dropping probability, we developed a simulation model. In our simulation, we consider ring network topology which consists of 16 core and 16 edge routers (nodes) with bidirectional links. Average node degree is 2 and average hop (H) is 5.8. A bidirectional link is realized by two unidirectional links in opposite direction. Each unidirectional link consists of 8 data channel and 1 control channel. Burst arrivals in the network are self-similar or poisson with arrival rate λ. Bursts are generated by using threshold-based assembly schemes and value of threshold is 40 KB. Packet length is kept as 2000 bytes. Shortest path is used for routing the burst from source to destination. Three different classes of traffic are considered namely class 1 (Lower priority), class 2 (Higher priority) and overall. Overall traffic is a combination of class 1 and class 2 traffic. CP processing time at each node (δ) is 1 μs. Offset time of class 1 traffic is δH. Offset time of class 2 traffic is 10δH which is greater than class 1 traffic. Range of traffic load is from 0.3 to 0.6. It is ensured that CP

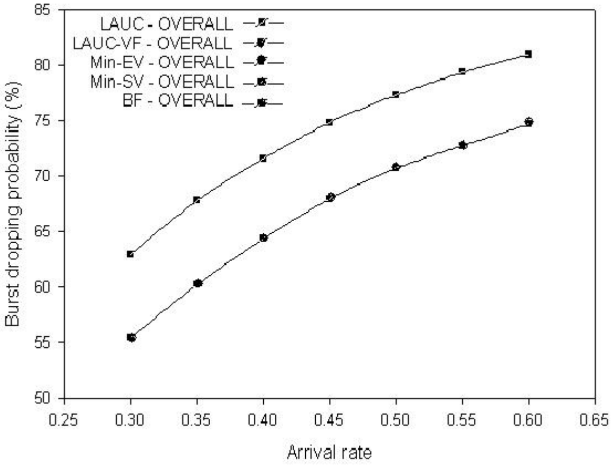


Fig. 5. Performance of overall traffic for various algorithms under self-similar traffic

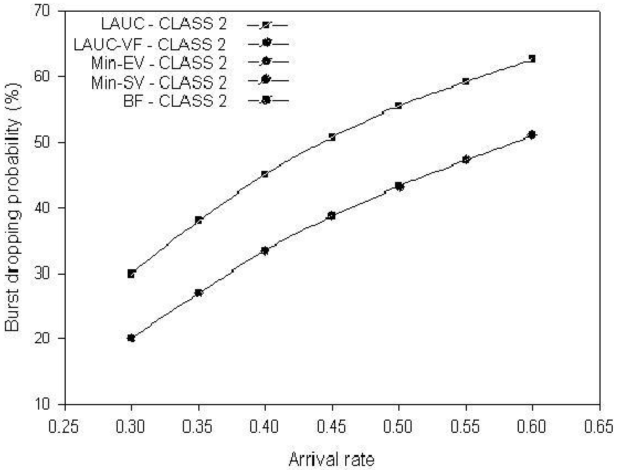


Fig. 6. Performance of class 2 traffic for various algorithms under self-similar traffic

is processed at each intermediate node well before the data burst is transmitted. Bursts are uniformly distributed over all sender-receiver pairs.

Figs. 5-7, indicates that void filling algorithms have better performance in terms of burst dropping probability than LAUC algorithm for overall, class 1 and class 2 traffic under self-similar traffic, respectively. It has been observed that all void filling algorithms have the same burst dropping probability for overall, class 1 and class 2 traffic. It is also observed that overall traffic have more burst dropping probability than class 1 and class 2. As the offset time value of class 2 is higher therefore the burst dropping probability of class 2 is less, but class 1

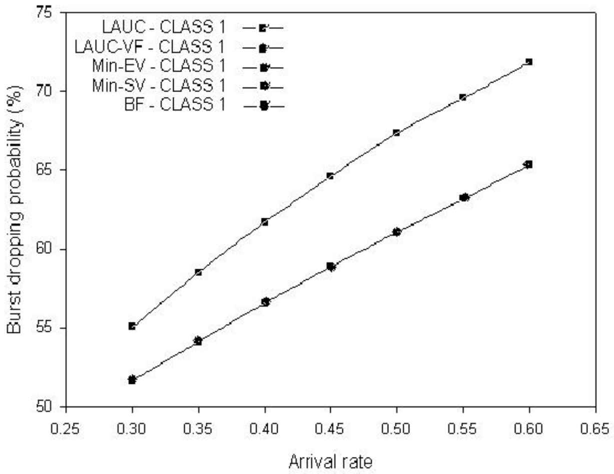


Fig. 7. Performance of class 1 traffic for various algorithms under self-similar traffic

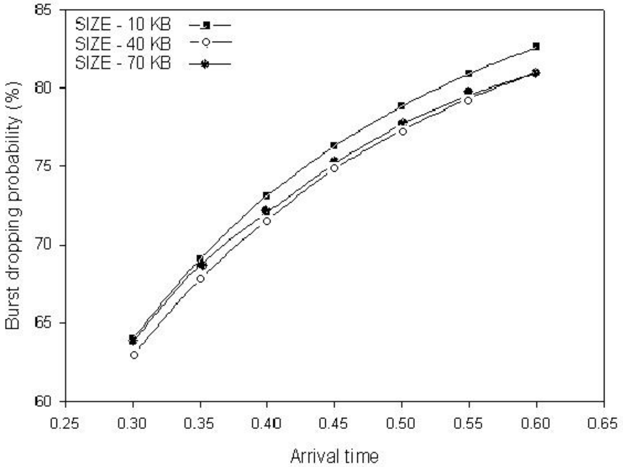


Fig. 8. Performance of overall traffic for LAUC algorithms under self-similar traffic with different threshold values

has smaller offset time value therefore the burst dropping probability of class 1 is more.

Figs. 8-9 depicts effectiveness of threshold size on void filling and non-void filling algorithms under self-similar traffic. It is observed that as the value of threshold size increased from 10 KB to 40 KB in the LAUC algorithm, burst dropping probability decreases. But as threshold size value increased from 40 KB to 70 KB, burst dropping probability increases. Fig. 9 shows that, in the

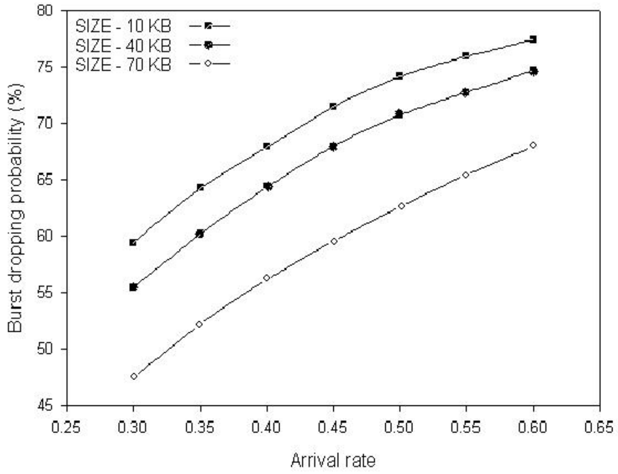


Fig. 9. Performance of overall traffic for various void filling algorithms under self-similar traffic with different threshold values

void filling algorithms, burst dropping probability is decreased with increase in the threshold size value from 10KB to 70KB.

Figs. 10-12 indicates that all void filling algorithm has better performance in terms of burst dropping probability than LAUC algorithm for overall, class 1 and class 2 traffic under poisson traffic, respectively. It is observed that, for

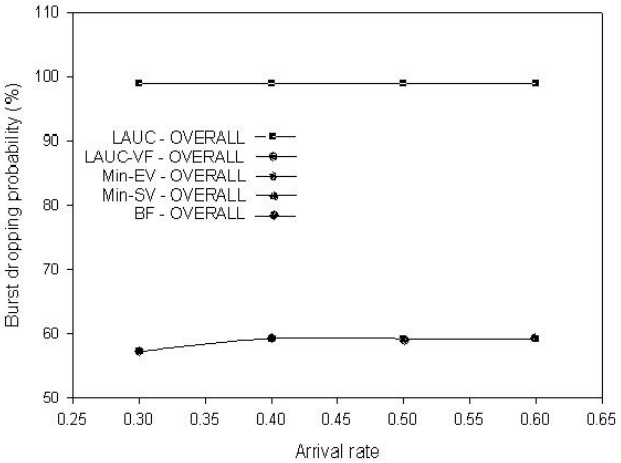


Fig. 10. Performance of overall traffic for various algorithms under poisson traffic

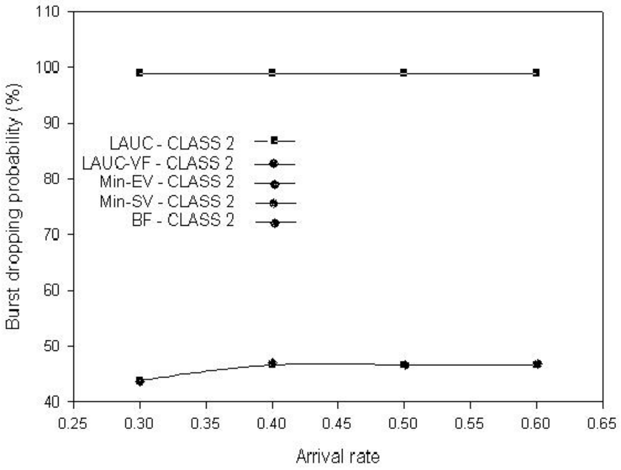


Fig. 11. Performance of class 2 traffic for various algorithms under poisson traffic

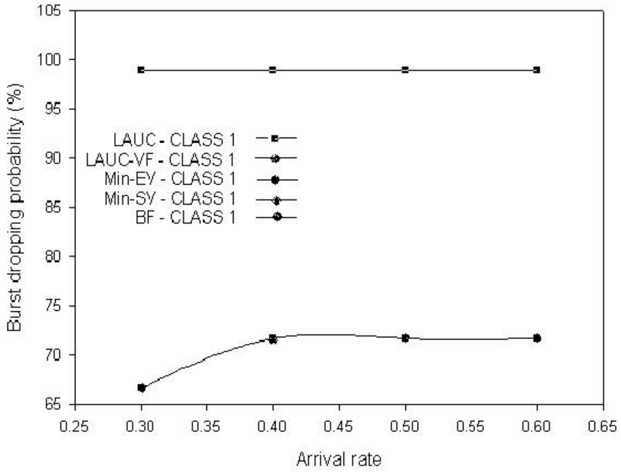


Fig. 12. Performance of class 1 traffic for various algorithms under poisson traffic

different traffic load, LAUC algorithm has same and constant burst dropping probability for overall, class 1 and class 2 traffic under poisson traffic.

Figs. 10-12 shows that all void filling algorithms has the same burst dropping probability for overall, class 1 and class 2 traffic under poisson traffic. It is also noticed that, for different traffic load, all void filling algorithms has the constant burst dropping probability for overall, class 1 and class 2 traffic. But in the case of void filling algorithms, class 1 traffic has more burst dropping probability than overall and class 2. As the offset time value of class 2 is higher therefore, the

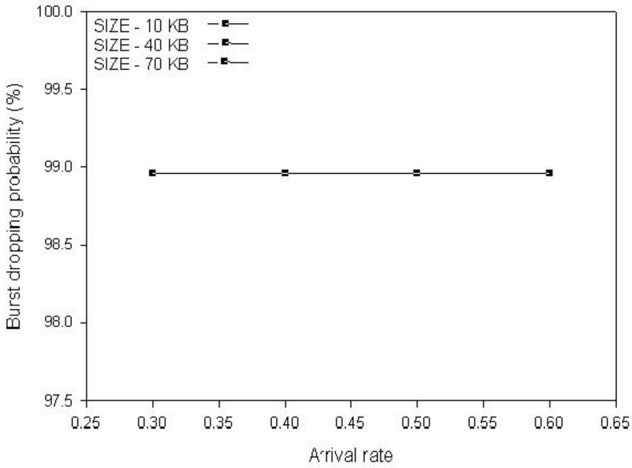


Fig. 13. Performance of overall traffic for LAUC algorithms under poisson traffic with different threshold values

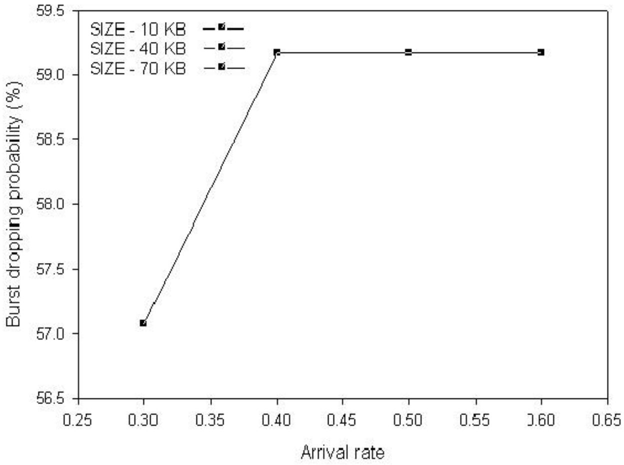


Fig. 14. Performance of overall traffic for various void filling algorithms under poisson traffic with different threshold values

burst dropping probability of class 2 is less. As class 1 has smaller offset time value therefore, the burst dropping probability of class 1 is more.

Fig. 13-14 depicts effectiveness of threshold size on void filling and non-void filling algorithms under poisson traffic. It is observed that, for different traffic load, there is no change in the burst dropping probability with increase in the threshold size.

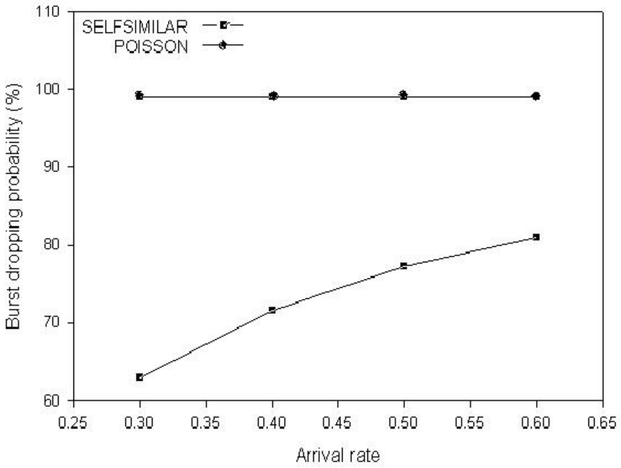


Fig. 15. Performance of overall traffic for LAUC algorithms under self-similar and poisson traffic

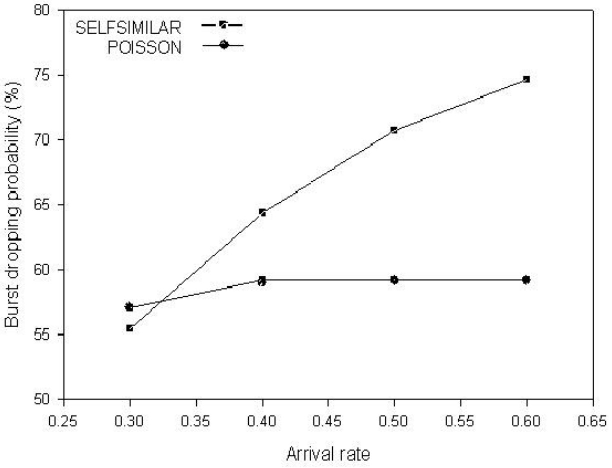


Fig. 16. Performance of overall traffic for various void filling algorithms under self-similar and poisson traffic

Fig. 15-16 compares the effectiveness of self-similar and poisson traffic on void filling and non-void filling algorithms. It is observed that poisson process suffers from large scale traffic analysis as burst dropping probabilities remain same and constant for different traffic load. While self-similar traffic does not suffers from large scale traffic analysis as burst dropping probability increases consistently with increase in traffic load.

4 Conclusion

In this work, under threshold burst assembly scheme with ring network, we studies the characteristics of all given factors which are affecting burst dropping probability. The different traffic like self-similar and poisson is applied to OBS network. The performance of LAUC, LAUC-VF, BF, Min-EV and Min-SV is investigated through simulation. It has been observed that threshold based scheme is poor scheme for burst generation. As threshold based scheme generate the same size of burst, all void filling algorithm like LAUC-VF, BF, Min-EV and Min-SV have the same burst dropping probability under self-similar and poisson traffic. It is also noticed that as the threshold value increases the burst dropping probability decreases. It has been noticed that the self-similar process gives us the better results with large scale analysis and generate huge amount of traffic but poisson process unable to provide it with same condition.

From the present results, our future work is to analyze the effect of hurst parameter and hopping count under self-similar traffic with threshold based scheme.

References

1. Mukharjee, B.: Optical WDM Networks. Springer, Heidelberg (2006)
2. Tzvetelina, B., Harry, P.: An Introduction to Optical Burst Switching. IEEE Optical Communication (2003)
3. Qiao, C., Yoo, M.: Optical Burst Switching: A New Paradigm for An Optical Internet. Journal of High Speed Network 8, 69–84 (1999)
4. Takuji, T., Soji, K.: Performance analysis of timer-based burst assembly with slotted scheduling for optical burst switching network. Performance Evaluation an International Journal 63, 1016–1031 (2006)
5. Yuhua, C., Pramod, K. V.: Secure Optical Burst Switching: Framework and Research Directions. IEEE Communication Magazine (2008)
6. Tzvetelina, B.: Optical Burst Switching: A Survey, Technical Report, NC State University, Computer Science Department (2002)
7. Chen, Y., Qiao, C., Yu, X.: An Optical Burst Switching: A New Area in Optical Networking Research. IEEE Netwoks 18, 16–23 (2005)
8. Praveen, B., Praveen, J., Siva Ram Murty, C.: A Survey of differentiated QoS schemes in optical burst switched networks. Science Direct Optical Switching and Networking 3, 134–142 (2006)
9. Jason, P., Jue, Vinod, V., Vokkarane: Optical Burst Switching Networks. In: Springer Publication, (2005)
10. Wang, R., Wu, D., Fang, G.: Data Burst Statictics and Performance Analysis of Optical Burst Switching Networks with Self-Similar Traffic. IEEE Computer Society, Los Alamitos (2007)
11. Burak, K., Oktug, S.F., Atmaca, T.: Performance of OBS techniques under self-similar traffic based on various burst assembly techniques. Computer Communications 30, 315–325 (2007)
12. Karamitsos, I., Varthis, E.: A Survey of Resrvation Schemes for OBS. In: University of Aegean, Department of information and Communication Sysytems,

13. Noureddine, B.: Optical burst switching protocols for supporting QoS and adaptive routing. Elsevier Computer Communications 26, 1804–1812 (2003)
14. Mohan, G., Akash, K., Ashish, M.: Efficient techniques for improved QoS performance in WDM optical burst switched networks. Elsevier Computer Communications 28, 754–764 (2005)
15. Konstantinos, C., Emmanouel, V., Kyriakos, V.: A new burst assembly scheme based on average packet delay and its performance for TCP traffic. Optical Switching and Networking 4, 200–212 (2007)
16. Jinhui, X., Chunming, Q., Jikai, L., Guang, X.: Efficient Channel Scheduling Algorithms in Optical Switched Networks using Geometric Technique. IEEE Journal on selected areas in Communication 22 (2004)
17. Jikai, L., Chunming, Q.: Schedule burst proactively for optical burst switching networks. Elsevier Computer Networks 44, 617–629 (2004)

An Optimization of Fundamental Frequency and Length of Syllables for Rule-Based Speech Synthesis

Kyawt Yin Win and Tomio Takara

Department of Information engineering, University of the Ryukyus, Okinawa, Japan
win@iip.ie.u-ryukyu.ac.jp, takara@ie.u-ryukyu.ac.jp

Abstract. In this paper an optimization method has been proposed to minimize the differences of fundamental frequency (F_0) and the differences of length among the speakers and the phonemes. Within tone languages use pitch variation to construct meaning of the words, we need to define the optimized fundamental F_0 and length to obtain the naturalness of synthetic sound. Large variability exists in the F_0 and the length uttered by deferent speakers and different syllables. Hence for speech synthesis normalization of F_0 and lengths are important to discriminate tones. Here, we implement tone rule by using two parameters; optimized F_0 and length. As an advantage in the proposed method, the optimized parameters can be separated to male and female group. The effectiveness of the proposed method is confirmed by the distribution of F_0 and length. Listening tests with high correct rates approve intelligibility of synthetic sound.

Keywords: Speech, Optimization, Normalization, Myanmar tone, Rule-based synthesis.

1 Introduction

There are some researches on optimal unit selection algorithm for corpus-based TTS system [1]. In our former research, we introduced Rule-based Myanmar speech synthesis system [2-3]. In that system fundamental speech units are demi-syllables with level tone. To construct the TTS system, monosyllabic words are analyzed and the parameters are obtained for synthesis of tones. Tone rules were F_0 linear pattern.

Within tone languages that use pitch variations to contrast meaning [4]. For example, Myanmar is a tonal language comprising four different lexical tones. Fig.1 shows an example of F_0 contour of the four Myanmar tones with the syllable /ma/ uttered by a male native speaker. Also Mandarin Chinese has four different lexical tones. The exact nature of the F_0 characteristics of Mandarin words is highly variable across utterances and speaker. Four lexical tones in isolated syllables can be characterized to mainly in terms of the shape of their F_0 contour. Therefore F_0 contour is the most crucial characteristic of tone. Furthermore duration of tones is also important [5]. Even rule-based speech synthetic system with linear F_0 pattern is very simple, it is important to define reliable value of F_0 and syllables length to implement synthesis rule. The acoustic of speech are notoriously variable across speakers. Large variability exists in the F_0 height and the length of syllables uttered by deferent

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 114–124, 2010.

speakers and different syllables [4]. Hence for speech synthesis optimization of F_0 and lengths are important and necessary to discriminate tones.

Standard Myanmar is used by 8 main races and sub-races as an official language. It is spoken in most of the country with slight regional variations. In addition, there are other regional variants that differ from standard Myanmar in pronunciation and vocabulary [6, 7, 8]. Accordingly a large variability exists in the F_0 and lengths among the speakers. Beside in Myanmar, however, tones are unique in their simplistic pattern not only related to F_0 but also more specifically and importantly in terms of length. Myanmar tones have different lengths between short-tone and long-tone groups. This is the basis for the proposed linear pattern for tone rule using optimized F_0 and optimized lengths.

In our former research, tone rule is implemented with linear pattern using the average F_0 and the averages of syllable's length which are normalized value. Even though, the reasonable high intelligibility of synthesized tone was confirmed through listening tests of synthesized words, there are some errors between male and female speech parameters.

In this paper we normalize F_0 and length of each tone, so that the square-sum of each difference between F_0 and its arithmetical average was minimized by using optimization method. The average F_0 for each word is selected from the frames at the center of syllable. The synthetic speeches are evaluated by listening tests. The results show that our proposed method gives high intelligibility of synthetic sound comparing with other tone synthesis rule with F_0 linear pattern, such as VieTTS [9].

The organization of the paper is as follows.

Section 1: Introduction
Section 2: Background of Speech Synthesis System
Section 3: Tone Synthesis Procedure using optimized F_0 and length
Section 4: Results and Discussion
Section 5: Conclusion.

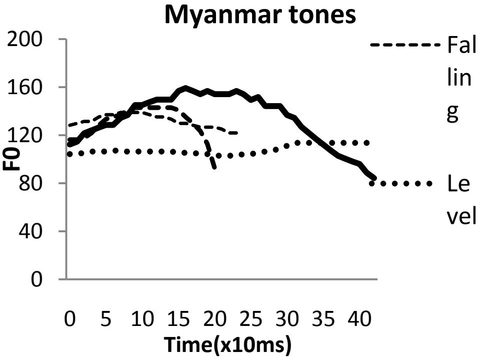


Fig. 1. Example of four tones of Myanmar syllable /ma/

2 Background of Speech Synthesis System

2.1 Speech Analysis and Synthesis

2.1.1 Speech Analysis

The analysis part of our speech synthesis system is designed using cepstral analysis. The frame length is 25.6ms and the frame shifting time is 10ms. As the window function for speech analysis, a time–domain Hamming window is used with a length of 25.6ms. The cepstral coefficient or cepstrum is defined as the inverse Fourier transform of the short-time logarithmic amplitude spectrum [10]. The special feature of the cepstrum is that it allows separating representation of the spectral envelope and excitation. The resulting parameter of speech units include the number of frames and, for each frame, voiced/unvoiced (V/UV) decision, pitch period and cepstral coefficients $c(m)$, $0 \leq \mathrm{m} \leq 29$.

2.1.2 Speech Synthesis

Under the control of the synthesis rule, the speech synthesis sub-system generates speech from pre-stored parameters. The source-filter model [11] is used as the speech production model. Fig.2 shows the structure of the speech synthesis sub-system in MyanmarTTS. The synthetic sound is produced using the Log Magnitude Approximation (LMA) filter, which has been introduced by Imai [12]. It presents the vocal tract characteristics. The spectral envelope is represented by the cepstral coefficients of 30 lower-order frequency elements. The LMA filter is a pole–zero filters that is able to represent efficiently the vocal tract features for all speech sounds.

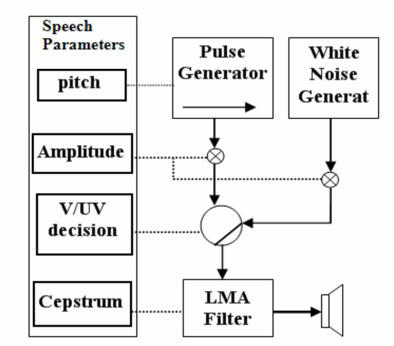


Fig. 2. MyanmarTTS speech synthesis sub-system

3 Tone Synthesis Procedure Using Optimized F_0 and Length

3.1 Tone Synthesis

The four Myanmar tones are analyzed to extract F_0 patterns. The data set is prepared as voiced sounds and meaningful words. We select consonant-vowel (CV) form with voiced consonants /b/, /m/, /l/ and three typical vowels /a/, /i/ and /u/. In total, 180 words (= 3 consonants x 3 vowels x 4 tones x 5 speakers) are used for tone analysis. After analyzing, four tones are distributed as shown in Fig.3. We find that the four tone groups overlapped and are not clearly discriminated. In our former research, we normalized F_0 and length to obtain relative values among the tones. The normalized parameters of tones using one syllable word were plotted in the distribution [3]. In this paper the normalized parameters by former normalization method using three syllables are shown in Fig.4.

3.2 Proposed Optimization Method

Lagrange's optimization method [13-14] is used for normalization. In this study we use 36 words of F_0 patterns by utterance of five native speakers. The words include three typical vowels "a", "i" and "u" with voiced consonants "b", "m" and "i". We select F_0 from three frames at the center of syllable word for each tone. The average F_0 values are selected from the middle frames of F_0 contours.

To minimize large differences of F_0 and differences of lengths among the speakers by means of tones, optimization method is carried out. The average of F_0 contours for each tone is given by

$$f_{ij}=1/n\sum_{k=1}^{n} f_{ij}^{k} \tag{1}$$

where n is number of F_0 contour. f_{ij} is F_0 at the center of syllable of i^{th} tone and j^{th} speaker.

Similarly, the average of tones is defined as A_j and the average of all speakers is defined as A.

To normalize $f_{ij,}$ Lagrange's optimization technique is utilized in this paper. For convenience, we define $U_{ij}{}^{0}$ and R_{ij} such as

$$R_{ij} = A\text{-}A_j \tag{2}$$

$$U_{ij}{}^{0}= f_{ij}{}^{0}\text{-}f_{ij} \tag{3}$$

where, $f_{ij}{}^{0}$ are normalized values of $f_{ij.}$

Then, in our problem, concentration of $f_{ij}{}^{0}$ around A is accomplished by minimizing

$$W(f_{ij}^{0}) = \sum_{j=1}^{s}(A-f_{ij}^{0})^2 \tag{4}$$

under the constraints

$$U_{ij}{}^{0} = \alpha_{ij}R_{ij} \tag{5}$$

where, α_{ij} are scale numbers and s is numbers of speaker.

Thus, normalized $f_{ij}{}^{0}$ are given by minimizing Lagrange's function L ($f_{ij}{}^{0}$)

$$\mathrm{L}\,(f_{ij}^{0}) = \mathrm{W}(f_{ij}^{0}) + \sum_{j=1}^{s} \lambda_j\,(U_{ij} - \alpha_{ij}R_{ij}) \tag{6}$$

For Eq. (6), we have

$$\frac{\partial \mathrm{L}}{\partial f_{ij}^{0}} = 2\left(f_{ij}^{0} - A\right) + \lambda_j \ = \ 0 \tag{7}$$

$$\frac{\partial \mathrm{L}}{\partial \lambda_j} = \ U_{ij} - \alpha_{ij} R_{ij} \ = \ 0 \tag{8}$$

Solving Eqs. (7), (8) gives

$$f_{\mathrm{ij}}^{0} = f_{ij} + \boldsymbol{\alpha}_{ij} R_{ij} \tag{9}$$

$$\lambda \mathrm{j} = 2(A- f_{ij} - \alpha_{ij} R_{ij}) \tag{10}$$

According to Eqs.(2) and (3), equation (5) indicates that if α_{ij} =1, f_{ij} around $A_{j,}$ i.e., f_{ij} - A_j is shifted to f_{ij}^{0} around A, i.e., f_{ij}^{0} –A, while $\alpha_{ij} = 0$, i.e., $f_{\mathrm{ij}}^{0} = f_{ij}$ which doesn't give normalization. When male and female speakers intermix, average A behaves as a center of A_j for male and A_j for female.

On the other hand, the minimum value of L is derived as follows:

$$\mathrm{L}_{\mathrm{min}} = \sum_{j=1}^{s} (A - f_{ij} - \alpha_{ij}^{0} R_{ij})^2 \tag{11}$$

which leads

$$\alpha_{ij}^{0} = (A- f_{ij}) / R_{ij} \tag{12}$$

because $\mathrm{L}_{\mathrm{min}} \geq 0$.

$$(A- f_{ij}) / R_{ij} > 0 \tag{13}$$

Hence, f_{ij} and Aj are always the same side of A.

Then, we have the relation

$$0 \leq \alpha_{ij} \leq \alpha_{ij}^{0} \tag{14}$$

From Eqs.(3) and (5),we get general equation

$$f_{\mathrm{ij}}^{0} = f_{ij} + \alpha_{ij} R_i \tag{15}$$

For the sake of convenience, we may simply choose α_{ij} in this paper, such that

$$\alpha_{ij} = \alpha = 1/2 \tag{16}$$

In this way f_{ij} is normalized. The normalized value f_{ij}^{0} is given by,

$$f_{\mathrm{ij}}^{0} = f_{ij} + \alpha R_{ij} \tag{17}$$

The optimized results are plotted in Fig. 5. Fig.5 (a), (b) show the distribution of four tones with optimized F_0 and optimized lengths, which are clearly discriminated in tone groups. From these figures we confirm that proposed method is an effective method to define the parameters for speech synthesis rule. Furthermore, as an advantage in the proposed method, the male and female can be distinguished.

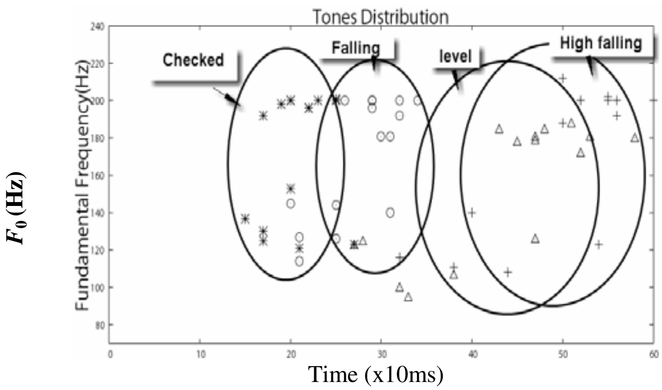


Fig. 3. Tones distribution of analysis-synthesis sounds by three female speakers and two male speakers before optimization

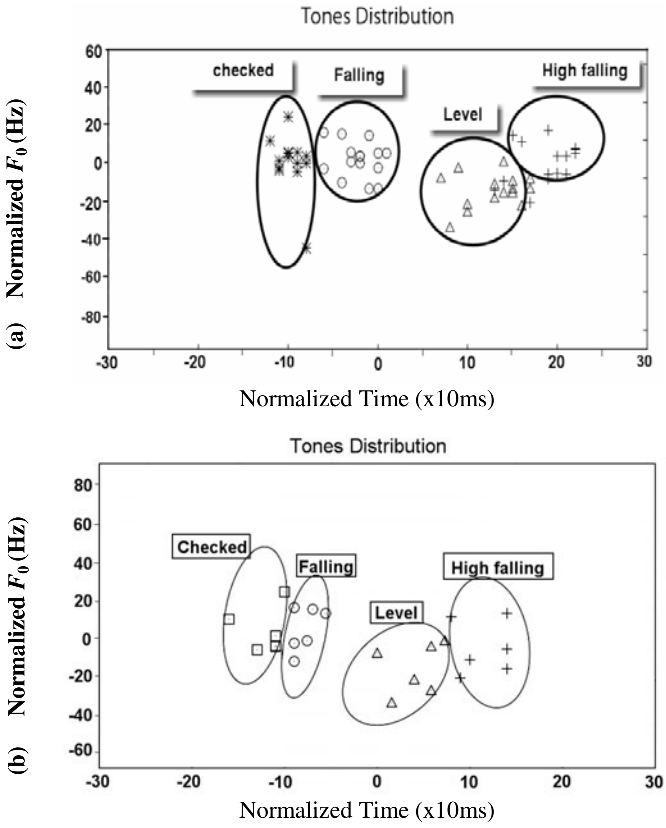


Fig. 4. (a - top) Tones distribution of analysis-synthesis sounds by three female speakers and two male speakers with normalized F_0 and normalized time (length). **(b – bottom)** Tones distribution of analysis-synthesis sounds uttered by two male speakers with normalized F_0 normalized time (length).

3.3 Tone Synthesis Rule with Linear F_0 Pattern

Myanmar tones are unique in their simplistic pattern not only related to F_0 but also more specifically and importantly in terms of length. Myanmar tones have different lengths between short-tone and long-tone groups. In accordance, after optimization we define tone rule employing two parameters; F_0 at the center of syllables and syllable's length as opposed to focusing on length alone. Tone rules are constructed with linear F_0 patterns.

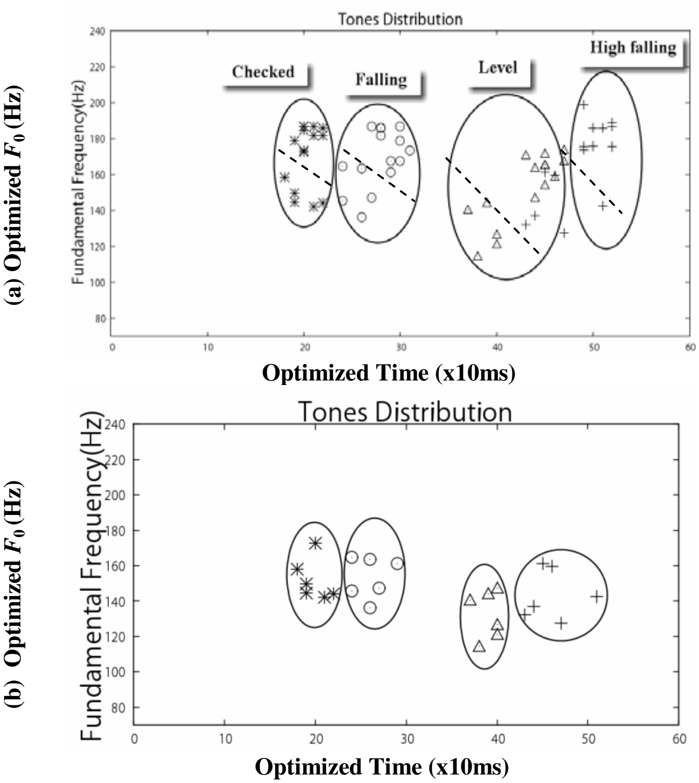


Fig. 5. (a - top) Tones distribution by three female speakers and two male speakers with optimized F_0, and optimized length. **(b - bottom)** Tones distribution by two male speakers with optimized F_0, and optimized length.

When we calculated the average frame length and average F_0 to make tone rules for male and female, we apply the concept of the center of gravity. As an example, Fig. 6 shows the calculation design of average F_0 and length using center of gravity. The tone rules are implemented based on optimized F_0 and optimized length of each tone as shown in Fig. 7.

We consider F_0 distribution as the mass distribution. We calculate average F_0 and length by using the concept of center of gravity x as follows:

$$x = (\sum_{i=1}^{n} x_i m_i)/M \tag{18}$$

$$M = m_1 + m_2 + m_3 + \ldots.. + m_n$$

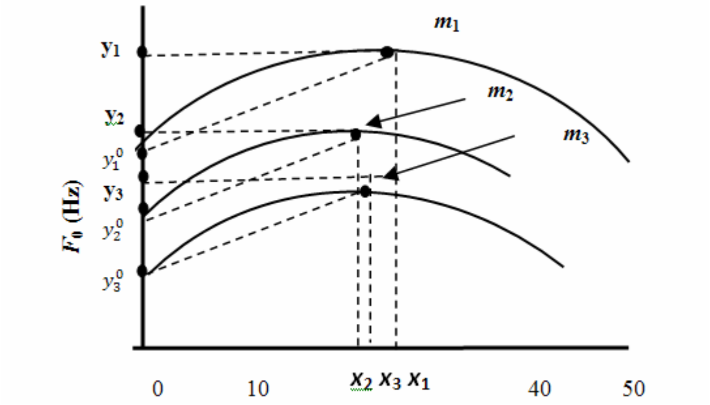


Fig. 6. The calculation design of average F_0 and length

Where m_i represents the weight of personal quality of F_0 of i^{th} speaker and x is average length of F_0 contour. Specifically, weight of personal quality of F_0 is different among the different speakers. As an example for three speakers,m_1,m_2 and m_3 are different values. In our experiments, all speakers are native and they have clear utterances and hearing ability. Therefore in this paper we consider their speech units have the same reliability. Then we have,

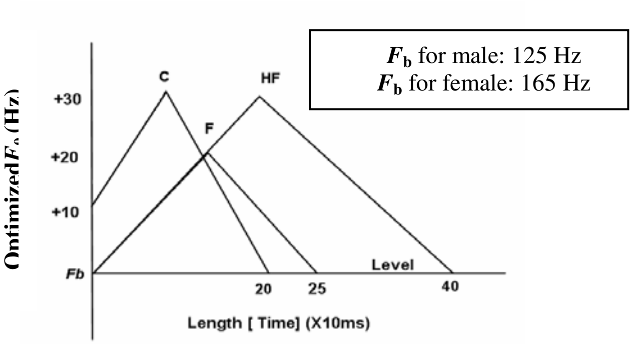
m_1=m_2=m_3= m (*Example*: for three speakers)

From Eq. (16) average F_0 value at the center of contour y is calculated as

$$y = \frac{m(y_1 + y_2 + y_3)}{3m} = \frac{(y_1 + y_2 + y_3)}{3} \tag{19}$$

Similarly the average length of time co-ordinate x is calculated as

$$x = \frac{(x_1 + x_2 + x_3)}{3} \tag{20}$$



L*: *Level tone, F*: *Falling tone, Hf*: *High falling tone, C: Checked tone,

Fig. 7. The diagram of tone rule

Using these rules, we carried out the listening tests to evaluate intelligibilities of synthetic speech of syllables and to evaluate the effect of proposed method.

4 Results and Discussion

Results of these tests are shown Fig. 8. These results have been obtained by using listening test. The result of our tone synthesis system and effectiveness of optimization are discussed as follows:

- Proposed method elicits the highest correct rate 99.68% for male speakers and 98.75% for female speakers.
- From these results we can confirm that optimized F_0 and length are conducted natural synthetic speech. Since we defined the scale factors of relative values properly, the optimized values are obtained.
- In VieTTS system[9], the result for linear pattern is about 85% for male, whereas the result of our system for male is 95.8%, even though our listening tests were done using the speech sounds of multiple speakers and different genders. Consequently, we can show that our linear pattern for tone rule is more effective than VieTTS's corresponding one since we applied the optimization method by means of multiple speakers and multiple phonemes.
- As a discussion concerning with above mentioned comparison, we consider that the optimization gives the effective values for both male and female, since we defined the scale factors of relative values correctly.
- Consequently, the introduced optimization method is effective and applicable for other speech synthesis rule for other tonal languages.

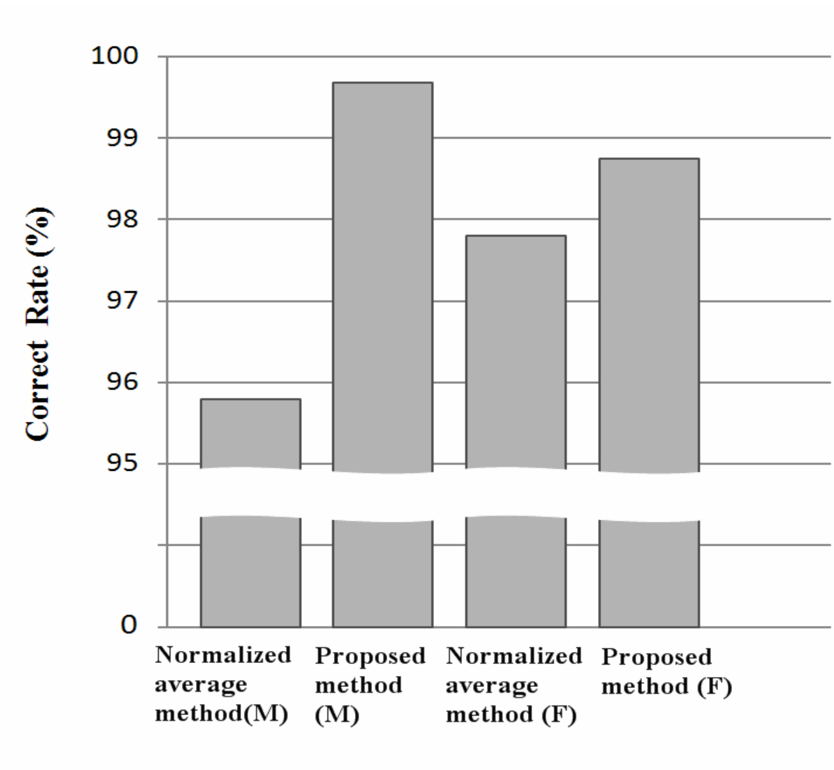


Fig. 8. The results of correct rate of perception of synthesized tone

5 Conclusion

An optimization method to define the parameters; F_0 and syllable's length for tone synthesis is introduced. We implemented tone rules of linear pattern based on two parameters, the optimized F_0 at the center of syllable and the optimized syllable's length. The effectiveness of the proposed method is confirmed by distribution of tones and the intelligibility scores of listening test. Although the high intelligibility of synthesized tone draws reasonably high correct rates in former research, the proposed method achieve the better results. Furthermore, in the proposed method, the optimized parameters can be separated into male and female groups. The introduced proposed method is applicable for other tone synthesis rule of other tonal languages.

References

1. Lee, M., Lopresti, D.P., Olive, J.P.: A Text-To-Speech Platform for Variable Length Optimal Unit Searching Using Perception Based Cost Function. International Journel Of Speech Technology 6, 347–365 (2003)
2. Win, K.Y., Takara, T.: Myanmar Speech Synthesis System Using Cepstral Method. In: The International Conference on Electrical Engineering (2008)

3. Win, K.Y., Takara, T.: Rule-based speech synthesis of Myanmar Using Cepstral Method. In: Proceeding of the 11th conference of Oriental-COCOSDA, NICT, Kyoto, Japan, November 25-27, pp. 225–229 (2008)
4. Huang, J., Holt, L.L.: General Perceptual Contributions to Lexical tone normalization. J. Acoust. Soc. Am. 125(6) (June 2009)
5. Zhang, S., Huang, T., Xu, B.: Tone Modeling for Contious Mandarin Speech Recognition. International Journel Of Speech Technology 7, 115–128 (2004)
6. Myanmar Language Committee, "Myanmar Grammar", Myanmar Language Committee, Ministry of Education, Myanmar (2005)
7. Thein Tun, U.: Some acoustic properties of tones in Burmese. In: Bradley, D. (ed.) Papers in South – East Asian Linguistics8: Tonation Canberra: Australian National University, pp. 77–116 (1982)
8. Wheatley, J.K.: Burmese. In: Cormier, B. (ed.) The World's Major Languages, pp. 834–845. Oxford University Press, New York
9. Do, T.T., Takara, T.: Vietnamese Text-To-Speech system with precise tone generation. Acoust. Sci. & Tech. 25(5), 347–353 (2004)
10. Noll, A.M.: Cepstrum Pitch Determination. Journal of the Acoustical Society of America 41(2), 293–309 (1967)
11. Furui, S.: Digital Speech Processing, Synthesis, and Recognition, 2nd edn., pp. 30–31. Marcel Dekker, Inc., New York (2001)
12. Imai, S.: Log Magnitude Approximation (LMA) Filter. Trans. IECE Jpn. J63-A, 886–893 (1980)
13. Xia, Y., Wang, J.: A General Methology for Desiging Globally Convergent Optimization Neural Networks. IEEE Transactions on Neural Networks 9(6) (November 1998)
14. Deng, L., Shaughnessy, D.O.: Speech Processing A dynamic and Optimization-Oriented Approach. Marcel Dekker, Inc., New York (2003)

Grid Smoothing for Image Enhancement

Guillaume Noel, Karim Djouani, and Yskandar Hamam

French South African Institute of Technology,
Tshwane University of Technology, Pretoria, South Africa

Abstract. The present paper focuses on sharpness enhancement and noise removal in two dimensional gray scale images. In the grid smoothing approach, the image is represented by a graph in which the nodes represent the pixels and the edges reflect the connectivity. A cost function is defined using the spatial coordinates of the nodes and the gray levels present in the image. The minimisation of the cost function leads to new spatial coordinates for each node. Using an adequate cost function, the grid is compressed in the regions with large gradient values and relaxed in the other regions. The result is a grid which fits accurately the objects in the image. In the presented framework, the noise in the initial image is removed using a mesh smoothing approach. The edges are then enhanced using the grid smoothing. If the level of noise is low, the grid smoothing is applied directly to the image. The mathematical framework of the method is introduced in the paper. The processing chain is tested on natural images.

Keywords: Grid smoothing, Sharpness enhancement, Mesh smoothing, Non-linear optimisation, Graph-based image.

1 Introduction

Image enhancement techniques seek at improving the appearance of an image without referring to a specific model for the degradation process, while *image restauration* relies on the knowledge of a degradation model [1]. The framework presented in the paper belongs to the *image enhancement* domain. The two main forms of image quality degradation are blur (loss of sharpness) and noise. Methods have been developed and address both types of degradation, either in a pixel-representation of the image or a mesh-representation. The pixel-representation considers an image to be a matrix of pixels. The edge enhancement methods modify the gray level of the pixels to improve the quality of the image. In a mesh representation, the image is represented by nodes (or vertices) and edges. Using the pixel-representation of an image, adaptive bilateral filters and quadratic-weighted median filters were applied with success for edge enhancement [1],[2]. These methods are based on filtering and may induce overshooting of the edges. Moreover, the parameters of the filters, in the case of the adaptive bilateral filter, are tuned according to a training dataset, narrowing the scalability to various applications. In any case, the performance of these methods is bounded by the use of the pixel-representation of the image. In low resolution images, the shape

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 125–135, 2010.

of the object does not systematically match the matrix and may lead to severe distortion of the original shape. For example, a clear straight line whose orientation is 45 degrees is represented by a staircase-like line. Image enhancement techniques, such as super-resolution ([3]) tackle the issue of the misrepresentation of an image. However, the square (even with a smaller size than the original pixel) is used as the construction brick of the image leading to the same misrepresentation of a shape ([3],[4] and [5]). The paradigm image and matrix may be overcome by the use of a mesh or graph-based representation of an image. The mesh defined on the image may be feature-sensitive or insensitive. In the feature-insensitive mesh, the definition of the mesh does not take into account the gray levels. The edge enhancement relies on the filtering of the gray levels in the image [6] which may lead to the same issues as in the pixel-representation (overshooting,...). In the feature-sensitive mesh, the graph is defined according to the distribution of the gray levels in the image and the position of the mesh is adjusted to the objects present in the image to enhance the image [7]. The usual problems of this class of techniques are the computation cost (interpolation) and the large number of edges in the graph. Mesh creation techniques may be found in [8], [9] and [10]. The present paper presents a novel combination of feature-insensitive and feature-sensitive mesh approaches, which reduces the complexity of the definition of the mesh and improves the representation of the information in the image. In the presented framework, an image is represented by a graph in which the nodes represent the pixels and the edges reflect the connectivity. The original graph (or grid) is a uniform grid composed by squares or triangles, depending on the connectivity chosen. The grid smoothing process modifies the coordinates of the nodes in the *(x,y)* plane while keeping the gray scale levels associated to the node unchanged. The grid smoothing relies on the minimisation of a cost function leading to a compression of the grid in the regions with large gradient values and a relaxation in the other regions. As a consequence, the new grid fits the objects in the image. The grid smoothing enhances the edges in the original image and does not modify the number of nodes. Noise removal techniques may be applied before the grid smoothing, depending on the properties of the original images. A new type of mesh smoothing is being used [11]. The mesh smoothing approach in [11] modifies the gray levels of the image while preserving the *(x,y)* coordinates of the nodes.

Section 2 of this paper presents the graph-based representation of an image while section 3 exposes the mathematical framework of the grid smoothing as well as the convergence properties. Noise removal using mesh smoothing is presented in section 4 and adapted to the grid smoothing context. Simulation results and examples of grid smoothing on real images may be found in Section 5. Conclusion and recommendations are underlined in section 6.

2 Graph-Based Image Representation

Our input data is a graph $G = (V, E)$, embedded in the 3D Euclidian space. Each edge e in E is an ordered pair (s, r) of vertices, where s (resp. r) is the sending (resp. receiving) end vertex of e [11]. To each vertex v is associated a triplet of real coordinates x_v, y_v, z_v. Let C_{ve} be the node-edge incidence matrix of the graph G, defined as:

$$C_{ve} = \begin{cases} 1 \text{ if } v \text{ is the sending end of edge } e \\ -1 \text{ if } v \text{ is the receiving end of edge } e \\ 0 \text{ otherwise} \end{cases} \quad (1)$$

In the rest of the paper, the node-edge matrix C_{ve} will also be denoted C.

Table 1. Number of connections according to the connectivity

Number of points	Number of connections Connectivity 4	Number of connections Connectivity 8
2500	4900	9702
10000	19800	39402
90000	179400	358202

Table 2. Computation time according to the connectivity ($\theta = 0.005$)

	Computation time (s)			
	Connectivity 4		Connectivity 8	
	Per edge	Image	Per edge	Image
2500 points	1.2×10^{-5}	5.9×10^{-2}	1.5×10^{-5}	1.5×10^{-1}
10000 points	1.6×10^{-5}	3.3×10^{-1}	2.2×10^{-5}	8.9×10^{-1}
90000 points	2.8×10^{-5}	5.0	3.3×10^{-5}	1.2×10^{1}

Considering an image with M pixels, X, Y and Z repectively represent $[x_1, ..., x_M]^t$, $[y_1, ..., y_M]^t$ and $[z_1, ..., z_M]^t$. X and Y are at first uniformely distributed (coordinates of the pixels in the plane), while Z represents the gray level of the pixels. Each pixel in the image is numbered according to its column and then its rows. We define L as the number of edges in the graph. C is consequently a matrice with L rows and M columns.

The number of edges depends on the choice of the connectivity for the pixel. If the connectivity is equal to four, each pixel is connected to its four closest pixels. The initial grid is then composed of squares. If the connectivity is equal to eight, each pixel is connected to its eight closest pixels. The initial grid is then composed of triangles. The choice of the connectivity is important as it increases the size of the matrix C and consequently the computation time required for the grid smoothing (Table 1,2). An evaluation of L may be derived using the dimensions of the image. If L_x and L_y represent respectively the number of pixels along the x axis and y axis, we have $M = L_x \times L_y$. For a connectivity equals to 4, $L = 2M - L_x - L_y$ and $L = 4M - 3L_x - 2L_y - 2$ if the connectivity equals 8. Using the notation introduced before, it may be observed that the complexity

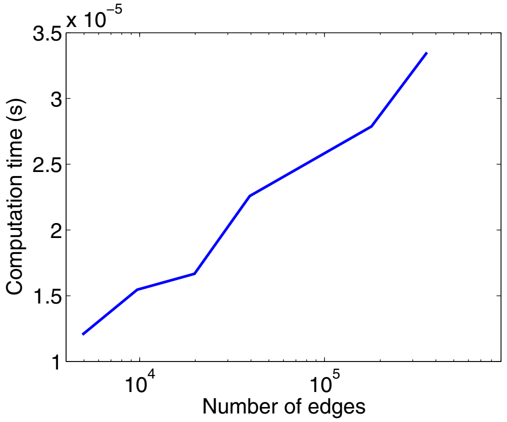
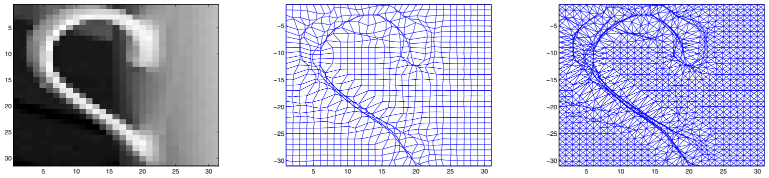


Fig. 1. Computation time per edge in seconds

of the algorithm is $L \times log(L)$(Fig.1). It may be explained by the complexity of the conjugate gradient with a stopping criterion ϵ in which the maximal number of iteration is bounded by $\alpha log(L/\epsilon)$, α being a constant. When using the high connectivity, the number of connections doubles as well as the computation time. The choice should be made according to the applications and the characteristics of the images. If an image includes thin lines which have to be preserved, the high connectivity should be used. For the other cases, the low connectivity gives satisfactory results.



(a) Original image (detail) (b) Grid (connectivity 4) (c) Grid (connectivity 8)

Fig. 2. Result of the grid smoothing according to the connectivity

3 Optimisation-Based Approach to Grid Smoothing

The present section introduces the framework for the grid smoothing. An extensive study of the convergence of the method as well as its application to satellite images may be found in [12], [14].

3.1 General Framework

A cost function is introduced to fit the content of the image with the grid. The main idea is that the regions where the variance is small (low gradient) require less points than the regions with a large variance (large gradient). The grid smoothing techniques will move the points of the grid from small variance regions to large variance regions. To achieve this goal, a cost function J is defined as follows:

$$J = J_X + J_Y \tag{2}$$

where

$$J_X = \frac{1}{2}\left[\left(X - \hat{X}\right)^t Q \left(X - \hat{X}\right) + \theta\left(X^t A X\right)\right] \tag{3}$$

and

$$J_Y = \frac{1}{2}\left[\left(Y - \hat{Y}\right)^t Q \left(Y - \hat{Y}\right) + \theta\left(Y^t A Y\right)\right] \tag{4}$$

with $A = C^t \varOmega C$ and $\hat{X}$ (resp. $\hat{Y}$) represents the initial coordinates of X (resp. Y). $\varOmega$ is a diagonal matrix. The first term in the expression of the cost function is called the *attachment* as it penalises the value of the cost function if the coordinates are too far from the original values. It is introduced to avoid large movement in the grid [11]. θ is a real number and acts as a weighing factor between the terms of the cost function. The matrix $\varOmega$ is defined as follows:

$$\varOmega_{k,k} = (z_i - z_j)^2 \tag{5}$$

where node i is the sending end of the vertex k and node j the receiving end. $\varOmega$ and Q are square diagonal matrices which dimensions are respectively $L \times L$ and $M \times M$.

As a result of the definition of $\varOmega$, the minimisation of J leads to the reduction of the areas of the triangle formed by two connected points and the projection of one of the point on the Z-axis. The edges in the image act as attractors for the points in the grid. As a consequence, the edges are better defined in terms of location and steepness in the smoothed grid.

3.2 Convergence of the Cost Function with Fixed Points and Attachment

The cost function with attachment results in a grid whose size might differ from the original grid size. A solution to conserve the original size is to fix the coordinates of the outer points of the grid. Let the X coordinates be partitioned into two parts, variable coordinates 'x' and fixed coordinates 'a' giving

$$X = \begin{bmatrix} x \\ a \end{bmatrix} \tag{6}$$

Then the first order cost function without attachment is

$$J_x = \frac{1}{2}\left(\left[\,(x - \hat{x})^t \; 0\,\right] Q \begin{bmatrix} (x - \hat{x}) \\ 0 \end{bmatrix} + \theta \left[\,x^t \; a^t\,\right] \begin{bmatrix} C_x^t \\ C_a^t \end{bmatrix} \varOmega \left[\,C_x \; C_a\,\right] \begin{bmatrix} x \\ a \end{bmatrix}\right) \tag{7}$$

Expanding the above equation gives

$$J_x = \frac{1}{2}\left[(x-\hat{x})^t Q_x (x-\hat{x}) + \theta x^t C_x^t \Omega C_x x + 2\theta x^t C_x^t \Omega C_a a + \theta a^t C_a^t \Omega C_a a\right] \quad (8)$$

The gradient of J_x with respect to x is

$$\nabla_x J_x = Q_x (x-\hat{x}) + \theta C_x^t \Omega C_x x + \theta C_x^t \Omega C_a a \quad (9)$$

Setting the gradient to zero gives

$$x = -\left[Q_x + \theta C_x^t \Omega C_x\right]^{-1}\left[Q_x \hat{x} - \theta C_x^t \Omega C_a a\right] \quad (10)$$

This gives the exact solution for the coordinates x.

Let x_{n+1} and x_n be x at iteration $n+1$ and n then

$$x_{n+1} = x_n - \alpha_n \nabla_x J_x \quad (11)$$

The gradient of J_x at the point x_{n+1} is equal to

$$\nabla_x J_{x_{n+1}} = \nabla_{x_n} J_x - \alpha_n Q_x \nabla_x J_{x_n} - \alpha_n \theta C_x^t \Omega C_x \nabla_x J_{x_n} \quad (12)$$

The optimal step condition may by expressed by $\nabla_x J_{x_n}^t . \nabla_x J_{x_{n+1}} = 0$

This leads to:

$$\alpha_n = \frac{\nabla J^t \nabla J}{\nabla J^t \left(Q_x + \theta C_x^t \Omega C_x\right) \nabla J} \quad (13)$$

The experience shows that the convergence is quicker using the conjugate gradient descent with optimal step. A quadratic function may be expressed by:

$$J(x) = \frac{1}{2} x^t A x + b^t x + c \quad (14)$$

where A is a definite positive matrix [13]. At each iteration, $x_{n+1} = x_n - \alpha_n d_n$, where α_n is the step and d_n is the direction of descent. The direction and the step are calculated at each iteration. By assimilation with the cost function with fixed points and attachment, we have $A = Q_x x + \theta C_x^t \Omega C_x$ and $b = \hat{x}^t Q_x - \theta a^t C_a^t \Omega C_x$. The step at the iteration n may be computed by:

$$\alpha_n = \frac{(b - Q_x \hat{x})^t (b - Q_x \hat{x})}{d_n^t Q_x d_n} \quad (15)$$

and the direction at iteration $n+1$ is equal to:

$$d_{n+1} = e_{n+1} + \frac{e_{n+1}^t e_{n+1}}{e_n^t e_n} d_n \quad (16)$$

where $e_{n+1} = e_n - \alpha_n Q_x d_n$ and $e_1 = b - Q_x \hat{x}$.

3.3 Stopping Criterion

As mentionned earlier, for large scale problem, the minimisation uses a gradient descent algorithm as it is computationally expensive to inverse very large matrices. Three gradient methods are used for the simulation, namely the steepest descent gradient with fixed step, the steepest descent gradient with optimal step and the conjugate gradient with optimal step. The descent gradient methods are iterative process and require a stopping criterion ϵ to stop the iterations. The chosen criterion is the simulation is the norm of the gradient ∇J. The iterative process continues while $\nabla J^t \nabla J \geq \epsilon$. When it is possible, the comparison between the exact coordinates given by the inversion of the matrix and the result of the gradient descent algorithm is small and is of the order of ϵ. For example, if $\epsilon = 10^{-3}$, the difference between the exact coordinates (matrix inversion) and the coordinates obtained through the gradient descent is 10^{-3} of the width of a pixel. The conjugate gradient descent is faster for any ϵ.

4 Noise Removal Using Mesh Smoothing

As mentionned in the sections above, the noise removal process is unrelated to the edge enhancement method in the grid smoothing framework. As there is no link between the two operations, the connectivity may be chosen differently. The following paragraph presents the noise removal process using a connectivity equal to four. The same operations may easily be derived for another connectivity. If the connectivity is equal to four, the grid on which the image is composed by quadrilaterals. Let's denote the quadrilateral Q_{ijkl}, the direct quadrilateral composed by the points i,j,k and l. Without noise removal, each quadrilateral represents a facet and the color of the facet is chosen equal to the graylevel of one point of the quadrilateral.

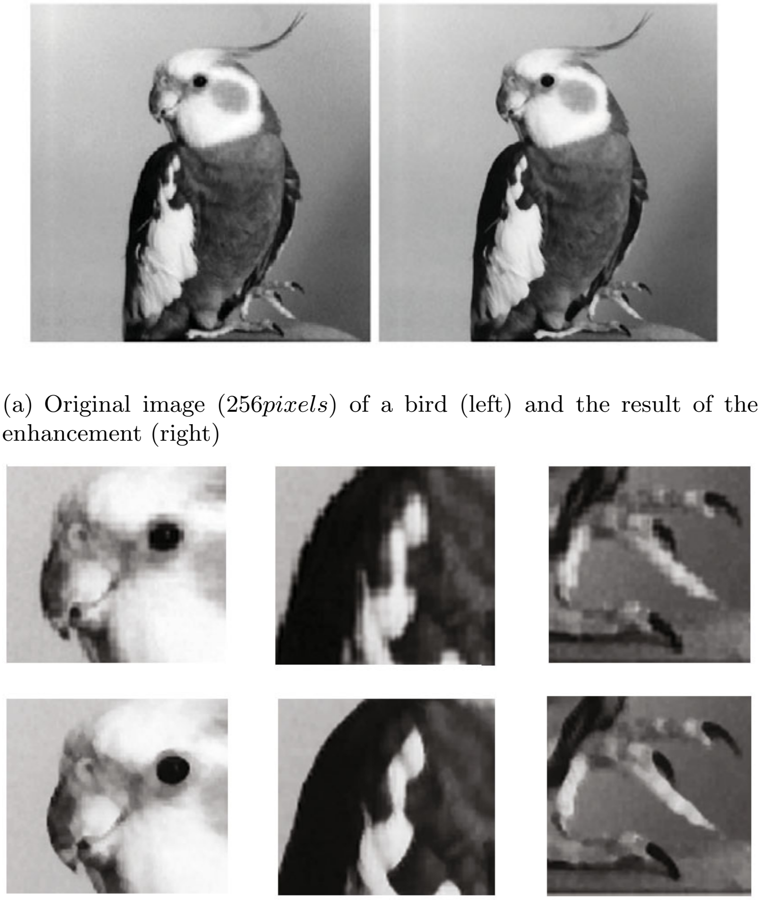
In the mesh smoothing framework, the image is represented using the matrix C_{ve} introduced in second section [11].The mesh smoothing techniques rely on the minimisation of a cost function J_Z, Z being the gray levels of the vertices. The result is a new vector $\bar{Z}$ containing the filtered gray levels of the image. The general form of the cost function is:

$$J_Z = \frac{1}{2}\left[\left(Z - \hat{Z}\right)^t Q \left(Z - \hat{Z}\right) + \theta_0 Z^t Z + \theta_1 Z^t \bar{A} Z + \theta_2 Z^t \bar{A}^2 Z\right] \qquad (17)$$

The cost function is composed of various terms whose respective weigths might be tuned by the values of the θ_i. The first term represents an attachment to the initial coordinates of the point to avoid large variations in graylevels [11]. The minimisation of the other terms of the cost function brings each point to the center of gravity of its neighbourhood. It may be shown that the second order term ($\theta_2 Z^t \bar{A}^2 Z$) smoothes the curve of the objects. For more information and the notations, please refer to [11]. The mesh smoothing used in the simulation takes into acount the attachment term and the second order term. It is refered as SOWA (Second Order With Attach) in [11].

5 Simulations

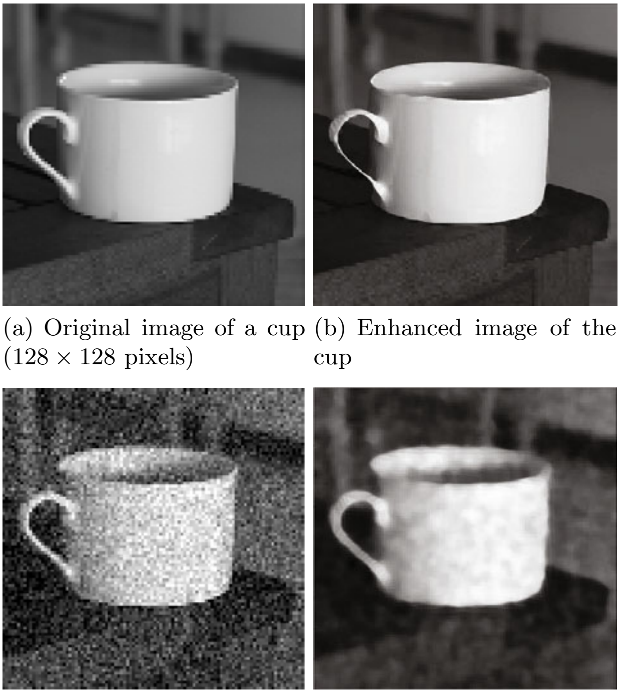
The simulations were performed using a standard laptop (1.87 GHz processor, 2GB RAM and *Windows Vista SP1* as operating system) and *Matlab R14 Service Pack 2*. Figure 2 shows the results of the grid smoothing process on a detail of the original image (Fig. 4(a)). Figure 2(b) shows the results of the grid smoothing with a connectivity is equal to 4 while Figure 2(c) presents the result when the connectivity is equal to 8. The results where obtained with a conjugate gradient descent. The regions with high variations in the graylevels needs more points than the other regions leading to a distorsion of the original grid. The



(a) Original image (256*pixels*) of a bird (left) and the result of the enhancement (right)

(b) Details of the bird image-Original(upper row) and enhanced (lower row)

Fig. 3.



(a) Original image of a cup (128 × 128 pixels)

(b) Enhanced image of the cup

(c) Original image with added noise

(d) Enhanced image with $\theta = 0.0005$

Fig. 4. Image restoration with noise removal

distorsions present in the two grids are similar. However, a higher connectivity leads to a more accurate and robust fitting of the shapes.

Figure 3 presents the results of the edge enhancement. The size of the original image (Figure 3(a)) is 256×256 pixels. It may be observed that the level of noise is low but that the edges are not well defined (pixelisation). The enhanced image (Figure 3(a)) exposes a good restauration on the edges while not compromising the quality of the image. The edges are smooth and continuous (the pixels which may be seen are due to the pdf compression of the image and are not present in the original simulation results). The texture of the bird is recovered while the dimension are slightly altered. A closer look at the improvment may be found in the details presented in Figure 3(b). Fig. 4 presents the results on a noisy version of the image(Fig. 4(a)). A white gaussian noise ($\mu = 0$ and $\sigma^2 = 0.1$) is added to the image. To avoid random movements of the points during the grid smoothing process, the image is firstly filtered using the mesh smoothing technique (SOWA). The filtered version of the image is then fed into the edge enhancement process. Fig. 4(d) depicts the results obtained with respectively $\theta = 0.0005$. It may be seen that the global shapes of the cup and the table are recovered in both cases. A greater θ leads to a better sharpening of the image while introducing distorsion in the shape represented. The choice of θ depends on the applications.

6 Conclusions

In conclusion, a new framework to enhanced images without a model for the degradation is presented in the paper. The method relies on the smoothing of the coordinates of the pixels in the image. The results of the image enhancement are promising. Combined with the mesh smoothing approach, the method performs well on noisy images. The output of the process is not a pixel-representation of the image. It leads to the main drawback of the technique which is the computational cost of the display of the facets. To overcome these limitations, further studies will involve redefinition of the connections in the grid to limit the number of facets. Another direction for future research will be to combine the mesh and the grid smoothing in a single operation, the objectif being to define a single cost function performing the same operations.

References

1. Zhang, B., Allebach, J.P.: Adaptative bilateral filter for sharpness enhancement and noise removal. IEEE Trans. on Image Processing 17(5), 664–678 (2008)
2. Aysal, T.C., Barner, K.E.: Quadratic weighted median filters for edge enhancement of noisy images. IEEE Trans. on Image Processing 13(5), 825–938 (2007)
3. Liyakathunisa Kumar, C.N.R., Ananthashayana, V.K.: Super Resolution Reconstruction of Compressed Low Resolution Images Using Wavelet Lifting Schemes. In: Second International Conference on Computer and Electrical Engineering ICCEE 2009, vol. 2, pp. 629–633 (2009)
4. Caramelo, F.J., Almeida, G., Mendes, L., Ferreira, N.C.: Study of an iterative super-resolution algorithm and its feasibility in high-resolution animal imaging with low-resolution SPECT cameras. In: Nuclear Science Symposium Conference Record NSS 2007, October 26-November 3, vol. 6, pp. 4452–4456. IEEE, Los Alamitos (2007)
5. Toyran, M., Kayran, A.H.: Super resolution image reconstruction from low resolution aliased images. In: IEEE 16th Signal Processing, Communication and Applications Conference, SIU 2008, April 20-22, pp. 1–5 (2008)
6. Wang, C.C.L.: Bilateral recovering of sharp edges on feature-insensitive sampled meshes. IEEE Trans. on Visualization and Computer Graphics 12(4), 629–639 (2006)
7. Xu, D., Adams, M.D.: An improved normal-meshed-based image coder. Can. J. Elect. Comput. Eng. 33(1) (Winter 2008)
8. Feijun, J., Shi, B.E.: The memristive grid outperforms the resistive grid for edge preserving smoothing, Circuit Theory and Design. In: Circuit Theory and Design ECCTD 2009, pp. 181–184 (2009)
9. Shuhui, B., Shiina, T., Yamakawa, M., Takizawa, H.: Adaptive dynamic grid interpolation: A robust, high-performance displacement smoothing filter for myocardial strain imaging. In: IEEE Ultrasonics Symposium, IUS 2008, November 2-5, pp. 753–756 (2008)

10. Huang, C.L., Chao-Yuen Hsu, C.Y.: A new motion compensation method for image sequence coding using hierarchical grid interpolation. IEEE Transactions on Circuits and Systems for Video Technology 4(1), 42–52 (1994)
11. Hamam, Y., Couprie, M.: An Optimisation-Based Approach to Mesh Smoothing: Reformulation and Extension. In: Torsello, A., Escolano, F., Brun, L. (eds.) GbRPR 2009. LNCS, vol. 5534, pp. 31–41. Springer, Heidelberg (2009)
12. Noel, G., Djouani, K., Hamam, Y.: Optimisation-based Image Grid Smoothing for Sea Surface Temperature Images. In: Advanced Concepts for Intelligent Vision Systems, ACIVS 2010, Sydney, Australia (2010)
13. Fletcher, R., Reeves, C.M.: Function Minimization by Conjugate Gradient. The Computer Journal, British Computer Society (1964)
14. Noel, G., Djouani, K., Hamam, Y.: Grid Smoothing: A graph-based Approach. In: 15th Iberoamerican Congress on Pattern Recognition, CIARP 2010, Sao Paulo, Brasil (2010)

Suppressing False Nagatives in Skin Segmentation

Roziati Zainuddin[1], Sinan Naji[1], and Jubair Al-Jaafar[2]

[1] Faculty of Computer Science and Information Technology, University of Malaya,
50603, Kuala Lumpur, Malaysia
roziati@um.edu.my, sinan@siswa.um.edu.my
[2] School of Computer Technology, Sunway University College,
64150, Kuala Lumpur, Malaysia
aljaafar@sunway.edu.my

Abstract. Human skin segmentation in colored images is closely related to face detection and recognition systems as preliminary required step. False negative errors degrade segmentation accuracy and therefore considered as critical problem in image segmentation. A general innovative approach for human skin segmentation that substantially suppresses false negative errors has been developed. This approach employed multi-skin models using HSV color space. Four skin color clustering models were used, namely: standard-skin model, shadow-skin model, light-skin model, and redness-skin model. The color information was used to segment skin-like regions by transforming the 3-D color space to 2-D subspace. A rule-based classifier produces four skin-maps layers. Each layer reflects its skin model. Pixel-based segmentation and region-based segmentation approaches has been combined to enhance the accuracy. The inspiring results obtained show that the suppression of false negatives is substantial and leads to better detection and recognition.

Keywords: Human skin segmentation, skin color modelling, face detection, HSV.

1 Introduction

Human skin segmentation in colored images is becoming an important task in many vision-based systems such as access control systems, face tracking, robot interaction, banking and financial transaction, video surveillance, videophone and teleconferencing, etc. The first task of such systems is to locate the face (or faces) within the image. It is not easy task since human face is a dynamic object and has a high degree of variability in its appearance (non-rigid object), which makes face detection a difficult problem in computer vision. Segmentation techniques based on color information as a cue has gained much attention motivated by four principle factors. First, color in general is a powerful descriptor that often simplifies object detection and extraction from a scene. Second, the processing of color information had proven to be much faster than processing of other facial features. Third, color information is robust against rotations, scaling and partial occlusions. Forth, Skin color information can be used as complimentary information to other features to improve detection. The challenges

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 136–144, 2010.

addressed with skin color segmentation can be attributed to the following factors: illumination, race, complex background, number of persons, imaging conditions, image montage, individual characteristics, aging, makeup, etc. Apparently these variations complicate skin segmentation and the larger the variations are, the more difficult the problem is.

Different skin color appearance caused by unconstrained scene conditions degrades segmentation accuracy. Segmentation may cause two kinds of errors: False Negative errors in which a skin pixel classified as a non-skin pixel, and False Positive errors in which an image pixel is declared to be skin pixel, but it is not. The most critical problem of the two errors is false negatives, attributed to the fact that, image segmentation is the first step in image analysis. Subsequently, when a face region is missed, the following stages of the system cannot retrieve the missed face. Therefore, false negatives determine the eventual success or failure of the subsequent stages.

The research aims to suppress false negative errors to achieve precise skin segmentation in fast, robust, and reliable approach. A novel approach has been introduced, so that, the approach is shifted from mono-skin model to multi-skin models using HSV color space. The detail description of building multi-skin color clustering models and the classification boundaries is presented in Section 3. The proposed approach for skin color segmentation is presented in section 4. Experimental results of the proposed approach and conclusion are given in Section 5 and 6 respectively.

2 Background

Numerous methods for skin segmentation and detection have been proposed so far. Each has its advantages and limitations. Some is superior to others whilst some yields the same result when compared to other technique.

A number of approaches have been proposed using different color spaces: RGB [1] [2], HSV or HSI [3] [4], YCbCr [5], YIQ [6], YES [7], CIE [8], YUV [9]. To build a skin color clustering model in the color space, many methods have been proposed. McKenna [10] has proposed Gaussian mixture models for the task which outperform single Gaussian model. Ruiz-del-Solar [2] has compensated for their color segmentation methods with additional features to obtain more valuable results robust to brightness variations. Gomez [11] has listed top components and made a hybrid color space from those. Jayaram [12] have proposed a method called an adaptive skin color filter that detects skin color regions in a color image by adaptively adjusting its threshold values. The Bayesian classifier with histogram technique has been used for skin detection by Jones [13]. Kim [4] proposed a skin color modeling approach in HSI color space while considering intensity information by adopting the B-spline curve fitting to make a mathematical model for statistical characteristics of a color with respect to intensity. Li [9] proposed an algorithm based on facial saliency map. Chen [1] proposed a hybrid-boost learning algorithm for multi-pose face detection and facial expression recognition. Juang [3] have used self-organizing Takagi–Sugeno-type fuzzy network with support vector is applied to skin color segmentation. Vezhnevets [14] wrote a survey on pixel-based skin color detection techniques.

A comprehensive survey on skin-color modeling and detection methods was written by Kakumanu [15].

3 Building Multi-skin Models

Generally, building skin model based on color information involves three main issues: First, what color space to choose. Second, how exactly the skin color distribution should be modeled, and finally, what will be the way of processing [14].

3.1 Choosing the Color Space

Algorithms based on skin color information need to deal with the sensitivity to the illumination conditions under which the input image is captured. HSV color space tends to be more realistic than other color systems. Its representation strongly relates to human perception of color, and it also eliminates the influence of illumination when describing the color of an object, Fig. 1(a). Hence, the HSV (and HSI) model is an ideal tool for developing image processing algorithms based on color descriptions [16]. The hue (H) is a measure of the spectral composition of a color and is represented as an angle from 0˚ to 360˚, while saturation (S) refers to the purity of a color and its value ranges from 0 to 1. Value component refers to the darkness of a color, which ranges also from 0 to 1. HSV color space had been chosen in our approach to build skin color models.

Generally, colored image contains millions of colors. The research also aims to reduce the number of colors in the source image. This will decrease the computational cost which is an essential to all systems. Reducing the number of colors process known as quantization. Each set of points of similar color is represented by a single color. Experimental results show that, clustering models of skin color are in the range of (0˚ ≤ Hue ≤ 49˚) and (339˚ ≤ Hue < 360˚) on the Hue wheel. Therefore, the range of colors at Hue wheel has been divided into equal intervals. Each interval step is (7˚). Hence, skin colors were reduced at Hue component to only eleven colors (Hue = 0, 7, 14, 21…, 49, and 339, 346, …, 360) which cover various types of skin; i.e. white, black, yellow, and redness skin colors under different lighting conditions.

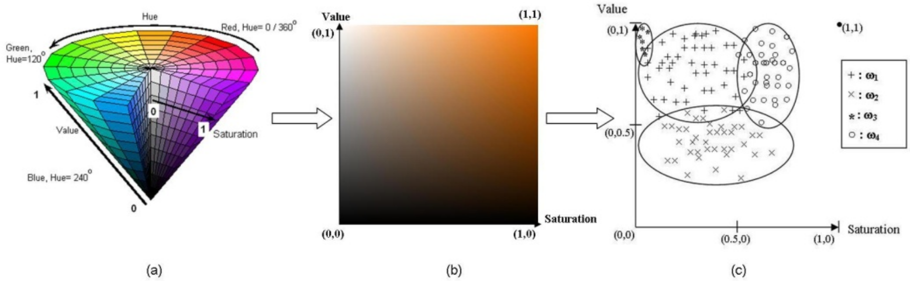


Fig. 1. Converting 3-D color space to 2-D subspace. (a) HSV color space. (b) 2-D S-V subspace, where Hue=28. (c) Four skin clusters of the pattern classes ω_1, ω_2, ω_3, and ω_4.

3.2 False Negative Errors Survey

The false negatives degrade segmentation performance dramatically with the diversity of image types and sources. To diagnose this problem, a survey on false negatives has been done. The survey showed that skin pixels classified as non-skin pixels can be categorized into four pattern classes:

1) Shadow regions (70%): shadows, blackish skin, and poor lighting.
2) Light regions (15%): because of strong light reflection, skin color information may be lost in some areas of the face.
3) Redness regions (8%): usually because of makeup, montage process, and flushing.
4) Others (7%)

The survey also reveals two main reasons behind the false negatives: first, the limitations of mono-skin model to cover many skin color appearance (dark, light, redness, etc). The second reason is due to the fact that each colored pixel is treated individually in relation to the color space (skin or non-skin pixel) without any consideration to the content of neighboring pixels.

3.3 Shifting from Mono-skin Model to Multi-skin Models

The avoidance of false negatives is clearly one of the functions of segmentation algorithm. One solution is to handle different skin color appearance caused by illumination variations and other factors in a proper way. This results into a novel approach that shifts from mono-skin model to multi-skin models. There is a good reason to use multi-skin models: pixels that are indistinguishable in one model may be fully distinguishable in the other model that will suppress the false negatives.

More than 3,500 face skin patches are used. These patches contain about (300,000) pixels of skin color data acquired from several regions of human skin including forehead, cheeks, nose, neck, and so forth. According to the skin color appearance, these patches have been divided into four pattern classes ω_1, ω_2, ω_3, and ω_4 namely:

ω_1: Standard skin color with uniform lighting
ω_2: Shadow skin, dark and blackish skin
ω_3: Light skin regions
ω_4: Redness skin

The next step involves the determination of optimum decision boundaries which are needed in the classification process.

3.4 Classification Rules

Determination of decision boundaries in three-dimensional color space is more difficult than in the two-dimensional. There is no easy or obvious way to enclose arbitrary clusters in three-dimensional space to see which pixels are selected for specific cluster. In most papers, low-dimensional color space is chosen instead of high-dimensional color space, to ease the determination process, (the R-G space replaces the R-G-B color space, the H-S space replaces the H-S-V color space, and so on). When the lighting is uniform, the segmentation performance is acceptable; but the performance is bad when dealing with unconstrained lighting conditions. This is due to loss of some information when an image is expressed in a low-dimensional space instead of a high-dimensional space. The research uses the full color information (three components: Hue, Saturation, and Value) in a novel way by transforming the 3-D color space to 2-D subspace without any color information losses. The idea is to use 2-D subspace for each constant quantized Hue value instead of 3-D color space. The corresponding transformed subspace would be S-V plane as shown in Fig. 1(b). The pattern vectors therefore, are of the form $\mathbf{x}=(s,v)'$ where s represents the Saturation component ($0 \leq s \leq 1$) and v represents the Value component ($0 \leq v \leq 1$) of the color features. The training samples of each pattern class tend to cluster about a typical representation within some region of the S-V subspace. Fig. 1(c) illustrate the four clusters in the S-V subspace where (Hue=28)[1]. The clusters are overlapped (non-separable situation) and not equally distributed.

A set of classification boundaries have been found using training algorithm. The algorithm is deduced from k-means algorithm and the reward-punishment concept. Here is an example of the classification boundaries that was obtained for Hue=28:

$P(x,y) \in \omega_1$ if $(0.12 < s(x,y) \leq 0.65)$ and $(0.6 \leq v(x,y) \leq 1)$
$P(x,y) \in \omega_2$ if $(0 \leq s(x,y) \leq 0.7)$ and $(0.33 \leq v(x,y) < 0.6)$
$P(x,y) \in \omega_3$ if $(0 \leq s(x,y) \leq 0.12)$ and $(0.75 \leq v(x,y) \leq 1)$
$P(x,y) \in \omega_4$ if $(0.65 < s(x,y) \leq 0.85)$ and $(0.6 \leq v(x,y) \leq 1)$

Where $s(x,y)$ and $v(x,y)$ represent the Saturation and Value components of the pixel at location (x,y). The pixel is defined as part of a specific pattern class when its two components s and v lie within the selected ranges. For different quantized Hue values, there will be different classification rules. These multiple rules are combined with Boolean OR operation. This is logically equivalent to segmenting each hue channel individually, creating separate binary images, and then combine them with Boolean OR operation afterward. It is clear that classification boundaries are more easily adjusted using 2-D subspace because of direct access to color space.

Classification rules in generally considered to be the high-end task of building skin color models. As shown in the experimental results section, the skin color models in this research can be used to detect the skin under different lighting conditions and under various types of skin color appearance.

4 Skin Color Segmentation

In general, finite level of accuracy can be achieved with pixel-based segmentation since each colored pixel is treated individually in relation to the color space (skin or non-skin pixel) without any consideration to the content of neighboring pixels. However, the skin of each human face has certain homogeneity between its pixels that could differentiate it from other objects. The approach in this research combines pixel-based segmentation and region-based segmentation to take in consideration the neighboring pixels that will improve segmentation accuracy. The input RGB image is

[1] For illustration, Hue=28 is found to be the most ideal case to graphically show the four skin clusters in Figure 1(c).

converted to the equivalent HSV color space. The number of colors in the input image is reduced by quantization. The input image is segmented using pixel-based segmentation approach which is based on classification rules. The classification rules produce four binary images called skin-maps in different separate layers (layer1, layer2, ..., layer4) as shown in Fig. 2(a) and Fig. 2(b). Fig. 2(b) shows that each layer reflects its relevant skin model. The merging process (region-based grow) between four skin-map layers is done to gather these pixels that satisfy the homogeneity between pixels of that region. The Region-based segmentation methods require the input of a number of seeds, might be manually, either individual seed pixels or regions. The approach in this research starts with the pixels in the first layer (ω1) as seed points. From these seed points, region grows by appending to the seed the neighboring pixels from the other layers (ω2, ω3, and ω4). Neighboring pixels are examined one at a time and added to the growing region if they are sufficiently similar. The conditions used:

- A pixel is adjacent to some seed pixel of the growing region,
- It belongs to some skin model at layers (ω_2, ω_3, and ω_4),
- Its color satisfies the similarity color of the growing region,
- It satisfies the texture analysis condition of the growing region and
- It is not an edge pixel.

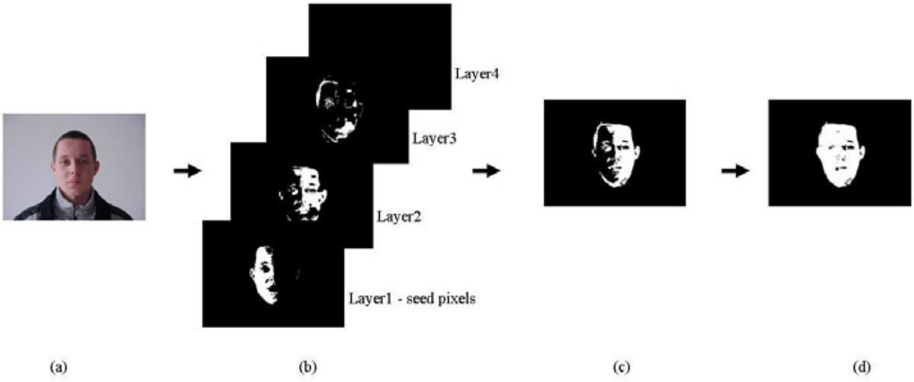


Fig. 2. Human skin segmentation. (a) Input image. (b) Pixel-based segmentation. (c) Region grows. (d) Skin segmentation output.

The color similarity condition used for growing region is not fixed; it is updated during merging process to reflect the actual mean intensity of the growing region to cover smooth changing in skin color. The Euclidian distance is used to measure color similarity:

$$\sqrt{(\mathrm{H_p} - \mathrm{H_{av}})^2 + (\mathrm{S_p} - \mathrm{S_{av}})^2 + (\mathrm{V_p} - \mathrm{V_{av}})^2} < \mathrm{T} \qquad (1)$$

where H_p, S_p, V_p are the intensity values of HSV components of the candidate pixel, and H_{av}, S_{av}, V_{av} are the average intensity value of HSV components of the growing region. The threshold of color intensity used is T, which is calculated locally.

The background objects at layer2, layer3, and layer4 that do not merge will be rejected at early stage of computation. Figure 2(c) shows the region grows output and Figure 2(d) shows skin segmentation output.

5 Experimental Results

A series of experiments were performed to evaluate the proposed approach, comparing the performance of mono-skin model approach to multi-skin models approach. For these experiments, three different databases were used. “The CVL Database" [17], which contains a total 114 people × 7 images with different views and expressions taken with uniform background. The “LFW database - Labeled Faces in the Wild” [18], is a database contains more than 13,000 images of faces collected from the Web. However, from the point of view of our objective, a dataset that contains 450 images were collected from the Web. In the experimentation part, 175 real images chosen carefully have been employed.

Table 1. System performance

Lighting Conditions	Using Mono-Skin Model		Using Multi-Skin Models	
	False Negatives Pixels	False Positives Pixels	False Negatives Pixels	False Positives Pixels
Uniform lighting	11%	7%	4%	9%
Un-uniform lighting (sidelight)	35%	6%	14%	10.5%
Gloomy scene	28%	8%	15%	13%

The images contain single face or multiple faces with various sizes and different lighting conditions. These images have been divided into three sets. The first set, represents the uniform lighting, the second set denotes lighting under sidelight, and the third set comprises images in gloomy scene. It is not a simple task to evaluate segmentation performance and compare algorithms. However, one of the most widely used criteria for performance evaluation is whether the system can outline the desired regions in the image. The criteria used for our system performance evaluation is as follows:

- The skin-maps for the test images have been manually annotated as a ground truth.
- Two approaches were applied; the conventional static mono-skin model compared to multi-skin models.
- Accordingly, two parameters for each system output have been measured relative to the ground-truth:
  (1) False Negative ratio (Rfn) is the number of falsely missed pixels to the actual number of skin pixels.
  (2) False Positive ratio (Rfp) is the number of falsely detected pixels to the actual number of skin pixels.

Table 1, shows a performance comparison between mono-skin model and multi-skin models. The segmentation performance of multi-skin models showed better performance. The important achievement is the suppression of false negatives that reveals its ability to segment skin regions with robustness to varying illumination conditions and different racial groups leading to better recognition in subsequent stages.

Fig. 3 clearly shows that the experiment results are promising. Although in Fig. 3 Row 1, the light was sidelight, the right part of the face has been correctly detected. A drawback of the approach is the slight increase in computational cost, which is mainly caused by the complexity of the improved multi-skin models. The second drawback of the approach is that false positives were slightly increased. However, this is not an issue since image segmentation, in general, is a preprocessing stage in face detection and recognition systems. It is obvious that color information on its own is not sufficient and that another stage based on other facial features is required to classify the skin segment to be a face or non-face. The experiments were carried out using PC Pentium 4 with a 3.00 GHz CPU and 1GB of RAM memory. The system was implemented in MATLAB (Ver. 7.9).

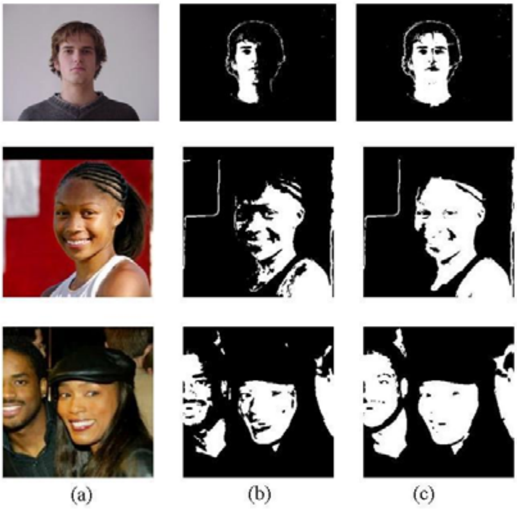


Fig. 3. Skin color segmentation. (a) Input image. (b) Using mono-skin model. (c) Using multi-skin models.

6 Conclusions

1- This approach based on using multi-skin models that substantially suppress false negative errors that reveals its ability to segment skin regions with robustness to varying illumination conditions and different racial groups leading to better detection and recognition in subsequent stages.
2- Higher suppression ratio can be achieved. However, it is undesired because the false positives will be slightly increased accordingly.

3- Multi-skin models caused higher computational cost than mono-skin model; however that is a normal outcome.
4- The approach does not require any preconditions and assumptions.
5- The approach can be applied to region segmentation for arbitrary objects (e.g., vehicles, hand tracking, aero imaging, etc).

References

1. Chen, W.C., Wang, M.S.: Region-Based and Content Adaptive Skin Detection in Color Images. International journal of pattern recognition and artificial intelligence 21(5), 831–853 (2007)
2. Ruiz-del-Solar, J., Verschae, R.: Skin Detection Using Neighborhood Information. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 463–468 (2004)
3. Juang, C.-F., Shiu, S.-J.: Using Self-Organizing Fuzzy Network With Support Vector Learning For Face Detection in Color Images. Neurocomputing 71(16-18) (2008)
4. Kim, C., You, B.-J., Jeong, M.-H., Kim, H.: Color Segmentation Robust to Brightness Variations by Using B-Spline Curve Modeling. Pattern Recognition 41(1), 22–37 (2008)
5. Chen, H.-Y., Huang, C.-L., Fu, C.-M.: Hybrid-Boost Learning for Multi-Pose Face Detection and Facial Expression Recognition. Pattern Recognition 41(3), 1173–1185 (2008)
6. Dai, Y., Nakano, Y.: Face-Texture Model Based on SGLD and Its Application in Face Detection in a Color Scene. Pattern Recognition 29(6), 1007–1017 (1996)
7. Saber, E.M., Tekalp, A.M.: Frontal-View Face Detection and Facial Feature Extraction Using Color, Shape and Symmetry Based Cost Functions. Pattern Recognition Letters 17(8), 669–680 (1998)
8. Kawato, S., Ohya, J.: Real-Time Detection of Nodding and Head-Shaking by Directly Detecting and Tracking The Between Eyes. In: IEEE International Conf. on Automatic Face and Gesture Recognition, Grenoble, pp. 40–45 (2000)
9. Li, B., Xue, X.Y., Fan, J.P.: A Robust Incremental Learning Framework for Accurate Skin Region Segmentation in Color Images. Pattern Recognition 40(12), 3621–3632 (2007)
10. McKenna, S., Gong, S., Raja, Y.: Modelling Facial Colour And Identity with Gaussian Mixtures. Pattern Recognition 31, 1883–1892 (1998)
11. Gomez, G.: On Selecting Colour Components For Skin Detection. In: International Conference on Pattern Recognition, vol. 2, pp. 961–964 (2002)
12. Jayaram, S., Schmugge, S., Shin, M.C., Tsap, L.V.: Effect of Color Space Transformation the Illuminance Component, and Color Modeling on Skin detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 813–818 (2004)
13. Jones, M.J., Rehg, J.M.: Statistical Color Models with Application to Skin Detection. Int. J. Computer Vision 46(1), 81–96 (2002)
14. Vezhnevets, V., Sazonov, V., Andreeva, A.: A Survey on Pixel-Based Skin Color Detection Techniques. Graphicon (2003)
15. Kakumanu, P., Makrogiannis, S., Bourbakis, N.: A Survey of Skin-Color Modeling and Detection Methods. Pattern Recognition 40, 1106–1122 (2007)
16. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Addison-Wesley, Reading (2002)
17. The CVL Database, `http://lrv.fri.uni-lj.si/facedb.html`
18. The LFW Database, `http://vis-www.cs.umass.edu/lfw/`

Steady State Analysis of an Energy Efficient Mobility Management Scheme in IP-Based Wireless Networks*

Ho Young Hwang[1], Sun-Jong Kwon[2], Yun Won Chung[3],
Dan Keun Sung[4], and Suwon Park[5,**]

[1] Department of Computer Engineering, Kwangwoon University, Seoul, Korea
hyhwang@kw.ac.kr
[2] KT, Seoul, Korea
[3] School of Electronic Engineering, Soongsil University, Seoul, Korea
[4] Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea
[5] Department of Electronics and Communications Engineering, Kwangwoon University, Seoul, Korea
spark@kw.ac.kr

Abstract. An energy efficient mobility management scheme in IP-based wireless networks was proposed to reduce the battery power consumption of mobile hosts (MHs). In order to efficiently manage battery power, radio resources, and network load, the proposed scheme controls six MH states: communicating, attention/cell-connected, attention/paging area-connected, idle, off/attached, and detached states. We derive the stationary probabilities and steady state probabilities of the MH states for the proposed scheme in IP-based wireless networks in compact form. The effects of input parameters on the steady state probabilities and the power consumption of MHs are investigated in the proposed scheme and conventional scheme based on Erlang and Gamma distributions. The analytical results provide guideline for proper selection of system parameters and can be utilized to analyze the performance of mobility management schemes in IP-based wireless networks.

Keywords: Energy efficient mobility management, steady state analysis, Mobile IP.

1 Introduction

Wireless networks have evolved toward an IP based network architecture. In such networks, mobility needs to be handled at the IP layer based on the Internet Engineering Task Force (IETF) concept. Many IP-based mobility protocols, such as Mobile IPv4 (MIPv4) [1], Mobile IPv6 (MIPv6) [2], and IP micro-mobility

* This work was supported by the Korea Research Foundation Grant funded by the Korean Government [KRF-2008-357-D00181].

** Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 145–153, 2010.

protocols (e.g., Cellular IP [3] and HAWAII [4]) have been proposed. MIPv4[1] and MIPv6[2] do not distinguish *idle* mobile hosts (MHs) from *active* ones. And the MIP protocols support registration, but not paging. Thus, a care-of-address (CoA) must be updated whenever an MH moves to a different subnet which is served by a different foreign agent (FA) in the MIPv4 [1] or by a different access router (AR) in the MIPv6 [2], regardless of the MH states, i.e., *active* and *idle*.

Various schemes on IP paging services for MHs have been proposed. The P-MIP [5][6] is an extension to MIP that is designed to reduce the signaling load in the core Internet. In the scheme, two MH states, i.e., *active* and *idle* were defined. In the *active* state, a registration occurs whenever an MH changes its cell. On the contrary, in the *idle* state, a registration occurs only if the MH changes its paging area. If there are any incoming data for the *idle* MH, paging is performed to find the exact location of the called MH.

Many enhancements for the MIP protocols, such as hierarchical Mobile IPv6 (HMIPv6) mobility management [7] and fast handover for Mobile IPv6 (FMIPv6) [8] have been proposed. The Network-based Localized Mobility Management (NETLMM) protocol [9][10] has been proposed by the IETF NETLMM working group recently. The IETF NETLMM working group has developed Proxy Mobile IPv6 (PMIPv6) [11]. In PIMPv6, the network supports IP mobility management on behalf of the MH. Kong *et al.* [12] investigated qualitative and quantitative comparisons between MIPv6 and PIMPv6. Zhou *et al.* [13] presented a PMIPv6-based global mobility management architecture and protocol procedure called GPMIP.

In our previous study [14], we proposed a mobility management scheme which considered the *detached* and *off* states in IP-based mobile networks, and analyzed an optimal rate of binding-lifetime-based registrations that results in minimum network cost when the registrations are utilized as a means of identifying the *off* MHs. In order to operate the MHs as a fully power-saving mode in the *idle* or *dormant* state, a mobility management scheme may need to distinguish the state for the paging area-based registration from the *idle* or *dormant* state.

In our previous study [15], we proposed an energy efficient mobility management scheme for reducing the power consumption of MHs in IP-based wireless networks. In order to efficiently manage battery power, radio resources, and network load, the proposed scheme controls six MH states: communicating, attention/cell-connected, attention/paging area-connected, idle, off/attached, and detached states. We analyzed the effects of parameters based on simple exponential distribution assumption on parameters.

In this paper, as an extension of our previous study, we derive the stationary probabilities and steady state probabilities of the MH states for the proposed scheme in IP-based wireless networks in compact form. The effects of input parameters on the steady state probabilities and the power consumption of MHs are investigated in the proposed scheme and conventional scheme based on more practical distributions such as Erlang and Gamma distributions on session holding time. The analytical results obtained by this paper provide guideline for proper selection of system parameters and can be utilized to analyze the performance of mobility management schemes in IP-based wireless networks.

2 An Energy Efficient Mobility Management Scheme in IP-Based Wireless Networks

An access router (AR), which acts as a default router to the currently served MHs, provides MHs with IP connectivity. In this paper, MIPv6 [2] is considered as a reference mobility protocol because it provides many advantages over MIPv4 [1]. In [15], an energy efficient mobility management scheme was proposed to control the following six states: communicating, attention/cell-connected, attention/PA-connected, idle, off/attached, and detached MH states. We note that the proposed scheme can be applied to the MIPv4 and MIPv6-based wireless networks. Communicating and attention/cell-connected MHs behave in the same manner as in MIP. Changes in the correspondent node (CN) and home agent (HA) are not required. Only minor changes in the MH and a paging agent (PAgnt) are needed. PAgnt performs paging related functionalities, and the PAgnt manages one or more paging areas. In a paging area (PA), there can be two or more ARs. A unique paging area identifier (PAI) can be used to establish the identity of the PA.

A communicating or attention/cell-connected MH registers its collocated care-of address (CCoA) at the corresponding HA as in MIP. Thus, the PAgnt is not involved in the MIP registration procedure. When a communicating or attention/cell-connected MH moves to a different cell which is served by a different AR, the MH performs a cell-based registration in the same manner as the MIP registration. It is assumed that an MH remains in the communicating state during a data session. When the data session is completed, the MH enters the attention/cell-connected state and an attention timer is reset and started. The attention timer is used to determine the instant when the MH enters the idle state. Upon expiration of an attention timer, an attentive MH which is in attention/cell-connected state or attention/PA-connected state enters the idle state by performing a PA-based registration. Through the PA-based registration, the MH registers a PAI of current PA at the PAgnt. At the corresponding HA of the MH, a paging agent care-of address (PAgnt-CoA) of current PAgnt is registered. When an idle MH moves to a different PA or PAgnt, the MH enters the attention/PA-connected state to perform the PA-based registration and the attention timer is reset and started. Data packets are tunneled to the PAgnt and buffered at the PAgnt when the data packets which are destined to an idle MH arrive at HA. Thus, the HA is unaware of the idle MH state. The paging request message from the PAgnt is sent to ARs in the paging area.

When an MH is switched off, the MH power-off state can be detected by a binding-lifetime-based registration and an unsuccessful paging. If HA or PAgnt sets a limitation on the maximum binding lifetime, the binding-lifetime-based registration can be used to detect the MH power-off state. If the network detects a silence for more than an agreed time period or if the MH does not respond to paging, the network considers the MH state to be detached.

3 Steady State Analysis of an Energy Efficient Mobility Management Scheme

We derive the stationary probabilities and steady state probabilities of the MH states for the proposed energy efficient mobility management scheme in IP-based wireless networks in compact form. MH state transitions are given by the diagram [15]. We analyze the MH state transitions by using a semi-Markov process approach because the residence time of the MH in each state is not exponentially distributed [16].

The stationary probabilities of the imbedded Markov chain are obtained by:

$$\pi_j = \sum_{i=1}^{6} \pi_i P_{ij}, \qquad j = 1, 2, ..., 6 \tag{1}$$

$$1 = \sum_{i=1}^{6} \pi_i, \tag{2}$$

where π_i is the stationary probability of state i, and P_{ij} is the state transition probability from states i to j. The state transition probability matrix, $P = [P_{ij}]$ for the MH state transitions is given by:

$$P = \begin{pmatrix} 0 & P_{12} & 0 & 0 & P_{15} & 0 \\ P_{21} & 0 & 0 & P_{24} & P_{25} & 0 \\ P_{31} & 0 & 0 & P_{34} & P_{35} & 0 \\ P_{41} & 0 & P_{43} & 0 & P_{45} & 0 \\ 0 & P_{52} & 0 & 0 & 0 & P_{56} \\ 0 & P_{62} & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{3}$$

And, from Eqns. (1) - (3) the stationary probabilities of the MH states are solved as follows:

$$\pi_1 = \frac{1}{D}[(P_{52} + P_{56}P_{62})\{P_{24}(P_{41} + P_{31}P_{43}) + P_{21}(1 - P_{34}P_{43})\}], \tag{4}$$

$$\pi_2 = \frac{1}{D}[(1 - P_{34}P_{43})(P_{52} + P_{56}P_{62})], \tag{5}$$

$$\pi_3 = \frac{1}{D}[P_{24}P_{43}(P_{52} + P_{56}P_{62})], \tag{6}$$

$$\pi_4 = \frac{1}{D}[P_{24}(P_{52} + P_{56}P_{62})], \tag{7}$$

$$\pi_5 = \frac{1}{D}[(1 - P_{12}P_{21})(1 - P_{34}P_{43}) - P_{12}P_{24}(P_{41} + P_{31}P_{43})], \tag{8}$$

$$\pi_6 = \frac{1}{D}[P_{56}\{(1 - P_{12}P_{21})(1 - P_{34}P_{43}) - P_{12}P_{24}(P_{41} + P_{31}P_{43})\}], \tag{9}$$

where $D = 1 + P_{52} + P_{56} + P_{56}P_{62} + P_{21}(P_{52} + P_{56}P_{62}) - P_{12}P_{21}(1 + P_{56}) + P_{24}(1 + P_{43})(P_{52} + P_{56}P_{62}) + P_{24}(P_{41} + P_{31}P_{43})(P_{52} + P_{56}P_{62}) - P_{12}P_{24}(1 + P_{56})(P_{41} + P_{31}P_{43}) - P_{34}P_{43}(1 + P_{56})(1 - P_{12}P_{21}) - P_{34}P_{43}(1 + P_{21})(P_{52} + P_{56}P_{62})$.

The steady state probabilities of the semi-Markov process are solved as:

$$P_1 = [1 - \frac{\lambda_s}{\lambda_s + \lambda_{\mathit{off}}} F_s^*(\lambda_{\mathit{off}})]^{-1} \frac{\mu_{\mathit{off}}}{\lambda_{\mathit{off}} + \mu_{\mathit{off}}} \frac{\lambda_s(1 - F_s^*(\lambda_{\mathit{off}}))}{\lambda_s + \lambda_{\mathit{off}}}, \tag{10}$$

$$P_2 = [1 - \frac{\lambda_s}{\lambda_s + \lambda_{\mathit{off}}} F_s^*(\lambda_{\mathit{off}})]^{-1} \frac{\lambda_{\mathit{off}}\mu_{\mathit{off}}}{\lambda_{\mathit{off}} + \mu_{\mathit{off}}} \frac{1 - e^{-\lambda_2}}{\lambda_1 + \lambda_c e^{-\lambda_2}}, \tag{11}$$

$$P_3 = [1 - \frac{\lambda_s}{\lambda_s + \lambda_{\mathit{off}}} F_s^*(\lambda_{\mathit{off}})]^{-1} \frac{\lambda_{\mathit{off}}\mu_{\mathit{off}}}{\lambda_{\mathit{off}} + \mu_{\mathit{off}}} \frac{(\lambda_1 + \lambda_c)(1 - e^{-\lambda_3})e^{-\lambda_2}}{\lambda_1(\lambda_1 + \lambda_c e^{-\lambda_2})} \cdot \frac{\lambda_1 F_r^*(\lambda_1 + \lambda_{PA}) + \lambda_{PA}}{\lambda_1[1 - e^{-\lambda_3} F_r^*(\lambda_1 + \lambda_{PA})] + \lambda_{PA}(1 - e^{-\lambda_3})}, \tag{12}$$

$$P_4 = [1 - \frac{\lambda_s}{\lambda_s + \lambda_{\mathit{off}}} F_s^*(\lambda_{\mathit{off}})]^{-1} \frac{\lambda_{\mathit{off}}\mu_{\mathit{off}}}{\lambda_{\mathit{off}} + \mu_{\mathit{off}}} \frac{(\lambda_1 + \lambda_c)e^{-\lambda_2}}{\lambda_1 + \lambda_c e^{-\lambda_2}} \cdot \frac{1 - F_r^*(\lambda_1 + \lambda_{PA})}{\lambda_1[1 - e^{-\lambda_3} F_r^*(\lambda_1 + \lambda_{PA})] + \lambda_{PA}(1 - e^{-\lambda_3})}, \tag{13}$$

$$P_5 = \frac{\lambda_{\mathit{off}}\mu_{\mathit{off}}}{\lambda_{\mathit{off}} + \mu_{\mathit{off}}} \frac{1}{\lambda_i + \mu_{\mathit{off}}} [1 - \frac{\lambda_r}{\lambda_i + \mu_{\mathit{off}}} (1 - F_r^*(\lambda_i + \mu_{\mathit{off}}))], \tag{14}$$

$$P_6 = \frac{\lambda_{\mathit{off}}}{\lambda_{\mathit{off}} + \mu_{\mathit{off}}} \frac{1}{\lambda_i + \mu_{\mathit{off}}} [\lambda_i + \frac{\lambda_r \mu_{\mathit{off}}}{\lambda_i + \mu_{\mathit{off}}} (1 - F_r^*(\lambda_i + \mu_{\mathit{off}}))], \tag{15}$$

where $\lambda_s = \lambda_i + \lambda_o$, $\lambda_1 = \lambda_i + \lambda_o + \lambda_{\mathit{off}}$, $\lambda_2 = (\lambda_1 + \lambda_c)T_A$, $\lambda_3 = \lambda_1 T_A$, λ_i is the mean arrival rate of incoming sessions, λ_o is the mean arrival rate of outgoing sessions, λ_{off} is the mean switch-off rate, $1/\mu_{\mathit{off}}$ is the mean switch-off duration, $1/\lambda_c$ is the mean cell residence duration, $1/\lambda_{PA}$ is the mean PA residence duration, and T_A is the attention timer value. The session holding time is assumed to be generally distributed with a density function $f_s(t)$ with a mean $1/\mu_s$, and $F_s^*(\theta)$ is the Laplace transform of $f_s(t)$. The interval of binding-lifetime-based registration is assumed to be generally distributed with a density function $f_r(t)$ with a mean of $1/\lambda_r$, and $F_r^*(\theta)$ is the Laplace transform of $f_r(t)$.

4 Numerical Examples

The effects of input parameters on the steady state probabilities and power consumption of the proposed scheme and conventional scheme are investigated. The values of input parameters with power consumption of the MH in state i, Pc_i assumed for numerical examples are shown in Table 1.

Fig. 1 shows the effect of mean session holding time $1/\mu_s$ on the steady state probabilities. It is assumed that an MH remains in the *communicating* state during the session holding time of a packet train transmission, i.e., a data session. It is more likely that MHs stay in the *communicating* state as the value of $1/\mu_s$ increases. On the contrary, it is less likely that MHs stay in the *idle* state. Thus, the probability P_1 increases, but the probability P_4 decreases as the value of $1/\mu_s$ increases.

Fig. 2 compares the power consumption of the MH for the proposed scheme with that for the conventional scheme with varying the values of μ_s for three

Table 1. Examples of the Input Parameters

Parameter	Value	Parameter	Value
λ_{off}	1/4 (/hour)	μ_{off}	1 (/hour)
λ_i	1 (/hour)	λ_o	1 (/hour)
μ_s	3600/300 (/hour)	λ_c	4 (/hour)
λ_{PA}	$\lambda_c/\sqrt{N_{cell/PA}}$ (/hour)	T_A	120/3600 (hour)
λ_r	1 (/hour)	$N_{cell/PA}$	16
Pc_1	1.00 (W)	Pc_2	0.80 (W)
Pc_3	0.70 (W)	Pc_4	0.01 (W)
Pc_5	0 (W)	Pc_6	0 (W)

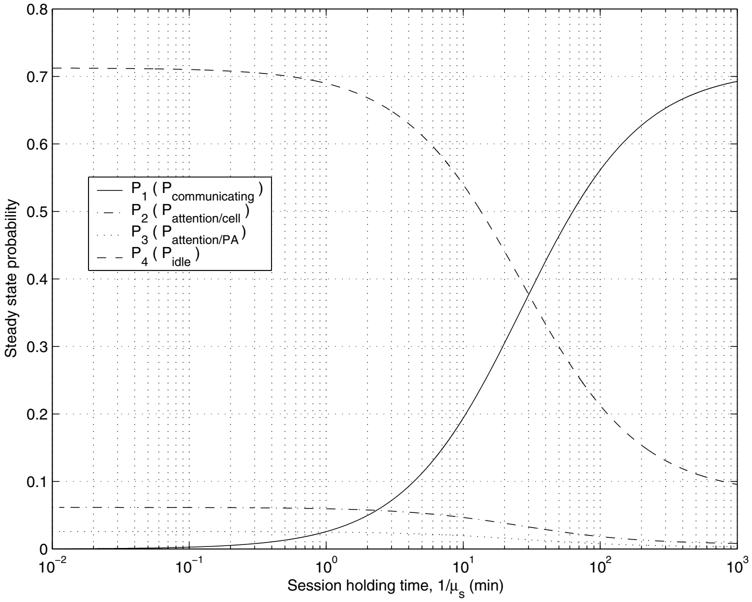


Fig. 1. The effect of μ_s on the steady state probabilities

types of distributions of session holding time: exponential (dotted line), Erlang (solid line), and Gamma (dashed line) distributions. The Gamma distribution can be used in approximating many other distributions. And it is useful for data packet transmission times because the Gamma distribution has the same trend as a Pareto distribution in terms of variance impact [17]-[20]. The Gamma distribution with shape parameter γ and scale parameter μ (i.e., mean $1/\mu_s = \gamma/\mu$ and a variance $v_{\mu_s} = \gamma/\mu^2$) has the following density function and Laplace transform:

$$f_s(t) = \frac{(\gamma\mu_s)^\gamma t^{\gamma-1}}{\Gamma(\gamma)} e^{-\gamma\mu_s t}, \qquad F_s^*(\theta) = \left(\frac{\gamma\mu_s}{\gamma\mu_s + \theta}\right)^\gamma, \qquad \gamma > 0. \qquad (16)$$

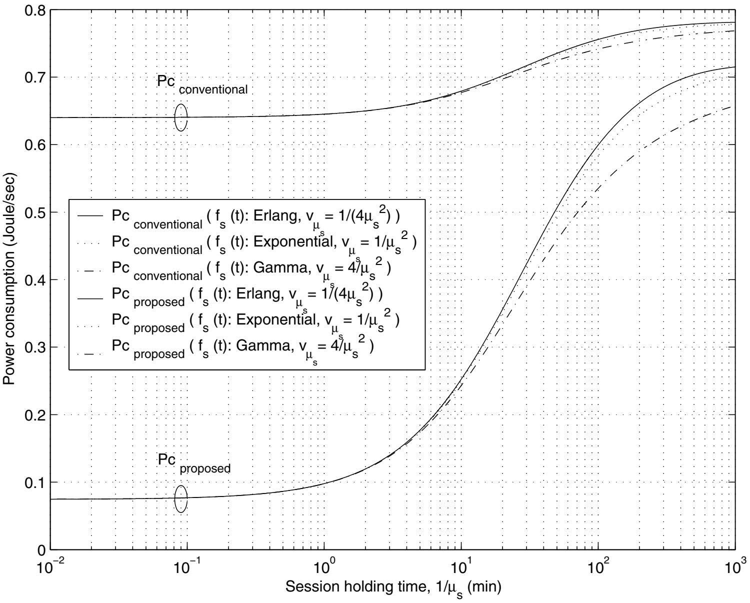


Fig. 2. The comparison of the power consumption in the proposed and conventional schemes for varying the values of μ_s when the session holding time is exponentially (dotted line), Erlang (solid line), and Gamma (dashed line) distributed

If γ is a positive integer, the Gamma distribution becomes an Erlang distribution. And if $\gamma = 1$, it becomes an exponential distribution. When the mean session holding time $1/\mu_s$ is lower than 1(min), the proposed scheme can save more than about 84.8%, compared with that of the conventional scheme for Mobile IP based wireless networks. If the mean session holding time $1/\mu_s$ is larger than 10 (min), the power consumption of the MH is sensitive to the distribution and variance of the session holding time $v_{\mu_s} = 1/(4\mu_s^2)$, $1/\mu_s^2$, and $4/\mu_s^2$ (i.e., $\gamma = 4$, 1, and 1/4).

5 Conclusion

An energy efficient mobility management scheme was proposed for reducing the power consumption of MHs in IP-based wireless networks. In order to efficiently manage battery power, radio resources, and network load, the proposed scheme controls the following six MH states: communicating, attention/cell-connected, attention/PA-connected, idle, off/attached, and detached states. We derived the stationary probabilities and steady state probabilities of the MH states for the proposed scheme in IP-based wireless networks in compact form. The effects of input parameters on the steady state probabilities and the power consumption of MHs are investigated in the proposed scheme and conventional scheme based on more practical distributions such as Erlang and Gamma distributions on session holding time. The proposed scheme yields a significant power saving in comparison with the conventional mobility management scheme for Mobile IP based wireless networks. The numerical examples show that the proposed scheme can save more than about 84.8% of the battery power consumption at the MH in comparison with the conventional scheme if the mean session holding time is low. According to the numerical examples, the power consumption of the MH is sensitive to the distribution and variance of the session holding time if the mean session holding time is large. The analytical results provide guideline for proper selection of system parameters and can be used to analyze the performance of mobility management schemes in IP-based wireless networks.

References

1. Perkins, C.: IP mobility support for IPv4. In: IETF RFC 3344 (August 2002)
2. Johnson, D.B., Perkins, C., Arkko, J.: Mobility support in IPv6. In: IETF RFC 3775 (June 2004)
3. Reinbold, P., Bonaventure, O.: IP micro-mobility protocols. IEEE Communications Surveys & Tutorials 5(1) 3rd qtr, 40–56 (2003)
4. Ramjee, R., Varadhan, K., Salgarelli, L., Thuel, S.R., Wang, S.-Y., La Porta, T.: HAWAII: a domain-based approach for supporting mobility in wide-area wireless networks. IEEE/ACM Transactions on Networking 10(3), 396–410 (2002)
5. Zhang, X., Gomez Castellanos, J., Campbell, A.T.: P-MIP: paging extensions for mobile IP. ACM/Springer Mobile Networks and Applications 7(2), 127–141 (2002)
6. Chung, Y.W., Sung, D.K., Hamid Aghvami, A.: Steady state analysis of P-MIP mobility management. IEEE Communications Letters 7(6), 278–280 (2003)
7. Soliman, H., Castelluccia, C., Malki, K.E., Bellier, L.: Hierarchical mobile IPv6 mobility management (HMIPv6). In: IETF RFC 4140 (August 2005)
8. Koodli, R.: Fast handover for mobile IPv6. In: IETF RFC 4068 (July 2005)
9. Kempf, J.: Problem statement for network-based localized mobility management (NETLMM). In: IETF RFC 4830 (April 2007)
10. Kempf, J.: Goals for network-based localized mobility management (NETLMM). In: IETF RFC 4831 (April 2007)
11. Gundavelli, S., Leung, K., Devarapalli, V., Chowdhury, K., Patil, B.: Proxy mobile IPv6. In: IETF RFC 5213 (August 2008)
12. Kong, K., Han, Y., Shin, M., Yoo, H., Lee, W.: Mobility management for all-IP mobile networks: mobile IPv6 vs. proxy mobile IPv6. IEEE Wireless Communications, 36–45 (April 2008)
13. Zhou, H., Zhang, H., Qin, Y., Wang, H.-C., Chao, H.-C.: A proxy mobile IPv6 based global mobility management architecture and protocol. ACM/Springer Mobile Networks and Applications 15(4), 530–542 (2010)
14. Kwon, S.-J., Nam, S.Y., Hwang, H.Y., Sung, D.K.: Analysis of a mobility management scheme considering battery power conservation in IP-based mobile networks. IEEE Transactions on Vehicular Technology 53(6), 1882–1890 (2004)
15. Hwang, H.Y., Kwon, S.-J., Chung, Y.W., Sung, D.K.: A mobility management scheme for reducing power consumption in IP-based wireless networks. In: Proc. Globecom 2002, vol. 3, pp. 2979–2983 (November 2002)
16. Ross, S.M.: Stochastic Process. John Wiley & Sons, Chichester (1996)

17. Lin, P., Lin, Y.-B.: Channel allocation for GPRS. IEEE Transactions on Vehicular Technology 50(2), 375–387 (2001)
18. Yang, S.-R.: Dynamic power saving mechanism for 3G UMTS system. ACM/Springer Mobile Networks and Applications 12(1), 5–14 (2007)
19. Lin, Y.-B.: Eliminating tromboning mobile call setup for international roaming users. IEEE Transactions on Wireless Communications 8(1), 320–325 (2009)
20. Lin, Y.-B., Tsai, M.-H., Dai, H.-W., Chen, Y.-K.: Bearer reservation with preemption for voice call continuity. IEEE Transactions on Wireless Communications 8(5), 2716–2725 (2009)

Performance Comparison of Handoff Modes in Terms of Energy*

Young-Uk Chung[1], Hyukjoon Lee[2], and Suwon Park[3]

[1] Department of Electronics Engineering, Kwangwoon University, Korea
[2] Department of Computer Engineering, Kwangwoon University, Korea
[3] Department of Electronics and Communications Engineering,
Kwangwoon University, Korea
{yuchung,hlee,spark}@kw.ac.kr

Abstract. Energy consumption becomes a critical issue within all industries. Especially, mobile communication systems is expected to be a major component of the world energy consumption budget. Handoff can highly affects the overall energy consumption of mobile communication systems. So it is important to taken into account energy consumption of handoff. In this paper, the transmission power to an MS during handoff is analyzed taking into account the macroscopic diversity. It is calculated for both hard handoff mode and soft handoff mode. The result shows that soft handoff mode is more efficient in view of energy consumption.

Keywords: Energy consumption, handoff, transmission power.

1 Introduction

Recently, the energy consumption becomes a critical issue within all industries. Many companies have been seeking methods to reduce energy consumption and operating costs actively. Especially, the industry about information and communication technology (ICT) is known as a major component of the world energy consumption budget. It was reported that the CO2 emission of the ICT industry exceeds the carbon output of the aviation industry [1]. It was estimated that mobile networks consumed nearly 80TWh electricity in 2008 [1] and the energy consumption of ICT is about 2% - 10% of the world energy consumption [2]. The main energy consumers in the ICT field are large data centers and servers, and wired and wireless communication networks. Among them, mobile communication systems are expected to contributes to global energy consumption even further in the future.

Therefore, communication system operators have been interested in energy saving approaches. Most interests focus on the radio access network (RAN) because large part of energy consumption occurs in the access network operations. Especially, power consumption in the base station (BS) is one of major sources. For example, in UMTS, one Node-B consumes around 1500 W on average and it

* This work is supported by the Research Grant of Kwangwoon University in 2010.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 154–163, 2010.

contributes about 60-80% of the whole network energy consumption [3],[4]. Due to their energy inefficiency and large quantities, BS sites are believed responsible for 80% of the energy consumption in mobile networks [5].

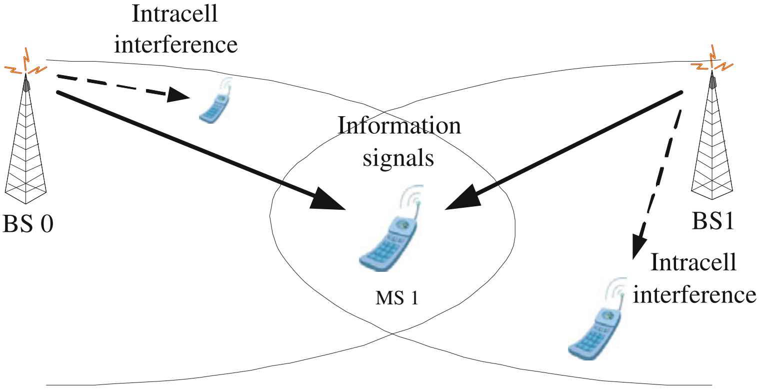


Fig. 1. Handoff Situation

Handoff is an inevitable process in cellular structured systems. There are two typical handoff modes: hard handoff and soft handoff. Hard handoff breaks off the initial connection with a BS before it switches to another BS. For this reason, it is referred to as "break-before-make" handoff. A brief break off before establishing connection with another BS is too short to be even noticed by users communicating on their mobile devices. On the other hands, soft handoff is a technique that makes mobiles residing in overlapping cell region communicate with both the currently serving BS and a target BS, simultaneously. In this case, the connection to the target BS is established before the connection to the source BS is broken. So, it is also called "make-before-break" handoff.

Though soft handoff provides better seamless property, it consumes more code and radio resource than hard handoff because soft handoff should remain two connections during handoff. It makes the shortage of system capacity. Explosive increasing market share of smartphone makes the use of data services over mobile networks increase dramatically. It has been reported that 3G data traffic volumes increase from 300% to 700% annually [6]. According to this trend, the capacity problem is expected to be even worse. Because, a mobile with data service consumes multiple radio resource simultaneously to support high speed transmissions. Therefore, high speed broadband mobile communication systems have adopted hard handoff as a default handoff mode.

However, it is also necessary to discuss which handoff mode is proper in view of energy consumption, besides this traditional viewpoints such as service quality and capacity. Because the effect of introducing a higher density of smaller cells in current systems causes frequent handoff, and the energy saved in one handoff procedure is multiplied by a large factor, with an important contribution to the reduction of the overall energy consumption in BS.

There has been various studies about handoff. However, most of these studies focus on the influence on system capacity and call admission control. No previous works have taken into account energy consumption of handoff. In this paper, the transmission power to an MS during handoff is calculated for both hard handoff mode and soft handoff mode. They are compared and discussed which mode is more efficient in view of energy consumption.

2 System Model

The cell model which we consider in this paper is shown in Fig. 2. We divide a cell into two regions such as normal region and handoff region. In normal region, a mobile communicates with only one BS. But a mobile is in soft handoff state when it resides in handoff region and communicates with both the serving BS and a target BS, simultaneously. For simplicity, we assume that a mobile during soft handoff can communicate with only two BS's, simultaneously. Normal region is surrounded by handoff region.

We assume that a system consists of K multiple cells which are uniformly distributed. Also, MS's are uniformly distributed in the system. Each MS receives J multipath rays from each BS. The rake receiver of MS is assumed to have $J^{'}(0)$ fingers. Traffic model is characterized by an activity factor that is defined

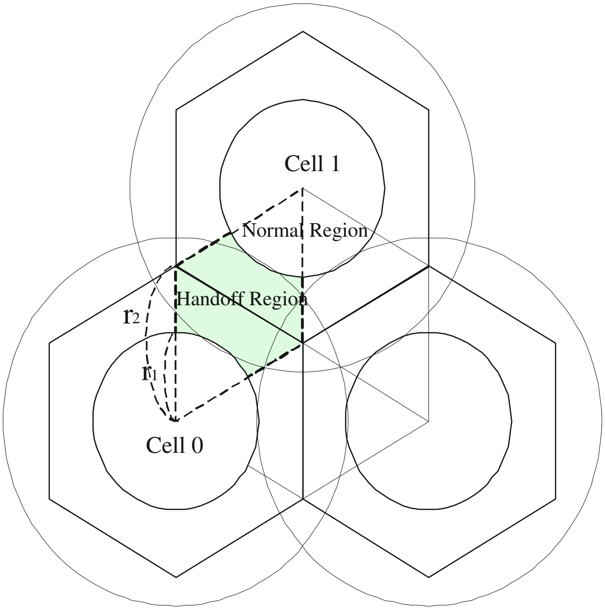


Fig. 2. Cell Model

as a duty cycle. The duty cycle is the percentage of time in which the user receives/transmits information during his call duration. From the traffic pattern, the activity factor of traffic can be obtained as

$$\rho = \frac{\sum_N T_{on}}{\sum_N (T_{on} + T_{off})} \approx \frac{E[T_{on}]}{E[T_{on}] + E[T_{off}]} \tag{1}$$

where N is the average number of sessions per a call, T_{on} is the duration of data transmission, and T_{off} is the time duration that there is no data transmission. T_{on} depends on the file size to be transmitted during the session and the transmission rate. We assume that each traffic source has constant data rate.

The path loss for a signal from $k-th$ BS to an MS is given by

$$L_k = 10^{\xi_k/10} \cdot r_k^{-4} = r_k^{-4} \cdot \chi_k \tag{2}$$

where r_k and ξ_k are a distance and a shadowing loss between the $k-th$ BS and the MS, respectively. ξ_k is Gaussian random variables with zero mean and σ standard deviation. χ_k is lognormal random variable. In general, σ has the value of $2-2.5$ dB for the signals from serving BS and $6-10$ dB for the signals from neighboring BS's when closed-loop fast power control is employed [7].

3 Analysis of Energy Consumption

In this section, we analyze the energy consumption of soft and hard handoff mode. We calculate the transmission power to an MS in handoff region for both soft and hard handoff mode. This analysis is extended based on the analytic model worked by Chung [8]. And we extended the analytic method into the environment taking into account the macroscopic diversity effect.

3.1 Soft Handoff Mode

Interference to an MS consists of intracell interference and intercell interference. Intracell interference occurs owing to signals from the serving BS. Because we assume that orthogonality in forward link is not guaranteed due to multipath, the signals which are transmitted from serving BS to the MS of interest through $j-th$ path are interfered by the signals which are transmitted to another MS's in the cell through multipaths except $j-th$ path. We define that P_k is total transmission power from $k-th$ BS. Then, the intracell interference for the $j-th$ path of an MS is given by

$$I^1_{intra} = (1-\alpha^j) \cdot (1-\psi \cdot \phi_i^{h0}) \cdot P_0 \cdot L_0 \tag{3}$$

$$I^2_{intra} = (1-\alpha^{J'(0)}) \cdot (1-\psi \cdot \phi_i^{h0}) \cdot P_0 \cdot L_0 \tag{4}$$

where α^j is power portion of $j-th$ path among multipaths, and ψ is a fraction of power assigned to traffic channels among total transmission power from a BS. ϕ_i^{h0} is also a fraction of power assigned to $i-th$ MS among total transmission power from $0-th$ BS, or serving BS.

Due to macroscopic diversity effect, the signals from target BS to an MS which is in soft handoff state do not interfere to the MS. In common, RAKE receiver of an MS receives the strongest one signal from target BS during soft handoff. So, the intercell interference of an MS which is located in handoff region is derived as

$$I_{inter}^1 = \sum_{k=1}^{K}\sum_{n=1}^{J} \alpha^n \cdot P_k \cdot L_k = \sum_{k=1}^{K} P_k \cdot L_k \tag{5}$$

$$I_{inter}^2 = \sum_{k=1}^{K} P_k \cdot L_k - \alpha^1 \cdot \psi \cdot \phi_i^{h1} \cdot P_1 \cdot L_1 \tag{6}$$

where ϕ_i^{h1} is also a fraction of power assigned to $i-th$ MS among total transmission power from 1^{st} BS, or target BS.

Let W be spreading bandwidth and R be data rate. Also we define ζ is combination gain according to each combining method. The total transmission power of a BS is determined by the number of on-going calls served by the BS. In this paper, we assume the worst situation that all BS's serve maximum number of on-going calls and transmit signals at full power level. And the interested MS is located at the cell boundary. Then, the E_b/N_t of the received signal at the MS is found as

$$\begin{aligned}(\frac{E_b}{N_t}) &= \frac{W}{R} \cdot \zeta \cdot \{ \sum_{j=1}^{J'(0)-1} \frac{\alpha^j \cdot \psi \cdot \phi_i^{h0} \cdot P_0 \cdot L_0}{I_{intra}^1 + I_{inter}^1 + N_0 \cdot W} + \frac{\alpha^1 \cdot \psi \cdot \phi_i^{h1} \cdot P_1 \cdot L_1}{I_{intra}^2 + I_{inter}^2 + N_0 \cdot W} \} \\ &\approx \frac{W \cdot \zeta \cdot \psi}{R} \cdot \{ \sum_{j=1}^{J'(0)-1} \frac{\alpha^j \cdot \phi_i^{h0} \cdot P_0 \cdot L_0}{I_{intra}^1 + I_{inter}^1} + \frac{\alpha^1 \cdot \phi_i^{h1} \cdot P_1 \cdot L_1}{I_{intra}^2 + I_{inter}^2} \}\end{aligned} \tag{7}$$

In this equation, we ignore $N_0 \cdot W$, because the background noise is negligible compared to the total transmission power.

For an MS in handoff region, we assume that the received signal powers from serving BS and target BS are approximately the same. Then, $P_0 \cdot L_0 \approx P_1 \cdot L_1 \approx P \cdot L$ and $\phi_i^{h0} \approx \phi_i^{h1} \approx \phi_i^h$. From this assumption, the E_b/N_t of the received signal at the MS can be given by

$$(\frac{E_b}{N_t}) \approx \frac{W \cdot \zeta \cdot \psi \cdot \phi_i^h \cdot P \cdot L}{R} \cdot \{ \sum_{j=1}^{J'(0)-1} \frac{\alpha^j}{I_{intra}^1 + I_{inter}^1} + \frac{\alpha^1}{I_{intra}^2 + I_{inter}^2} \} \tag{8}$$

Then, the relative power of $i-th$ MS in handoff region ϕ_i^h is derived as

$$\begin{aligned}\phi_i^h &= \frac{(E_b/N_t)\cdot R}{W\cdot\zeta\cdot\psi}\cdot\frac{1}{P\cdot L\cdot\{\sum_{j=1}^{J'(0)-1}\frac{\alpha^j}{I_{intra}^1+I_{inter}^1}+\frac{\alpha^1}{I_{intra}^2+I_{inter}^2}\}} \\ &= \frac{(E_b/N_t)\cdot R}{W\cdot\zeta\cdot\psi} \\ &\cdot\frac{1}{\sum_{j=1}^{J'(0)-1}\frac{\alpha^j}{(1-\alpha^j)\cdot(1-\psi\cdot\phi_i^h)+\sum_{k=1}^{K}\frac{L_k}{L}}+\frac{\alpha^1}{(1-\alpha^{J'(0)})\cdot(1-\psi\cdot\phi_i^h)-\alpha^1\cdot\psi\cdot\phi_i^h+\sum_{k=1}^{K}\frac{L_k}{L}}}\end{aligned} \tag{9}$$

Since $\psi\cdot\phi_i^h \ll 1$, the relative power for the $i-th$ MS can be approximated as

$$\phi_i^h \approx \frac{(E_b/N_t)\cdot R}{W\cdot\zeta\cdot\psi}\cdot\frac{1}{\sum_{j=1}^{J'(0)-1}\frac{\alpha^j}{(1-\alpha^j)+\sum_{k=1}^{K}\frac{L_k}{L}}+\frac{\alpha^1}{1+\sum_{k=1}^{K}\frac{L_k}{L}}} \tag{10}$$

Equation (10) which we derived is the relative power of an MS in the worst situation. For MS's in handoff region, the worst case means that the MS is located at the center of handoff region and serving BS and target BS transmit signals with the same power. Using this equation (10), we can calculate the average relative power of an MS in handoff region.

3.2 Hard Handoff Mode

The intracell interference of $j-th$ path of an MS in normal region is given by

$$\begin{aligned}I_{intra} &= \sum_{n=1,n\neq j}^{J}\alpha^n\cdot(1-\psi\cdot\phi_i^n)\cdot P_0\cdot L_0 \\ &= (1-\alpha^j)\cdot(1-\psi\cdot\phi_i^n)\cdot P_0\cdot L_0\end{aligned} \tag{11}$$

where α^j is power portion of $j-th$ path among multipaths, and ψ is a fraction of power assigned to traffic channels among total transmission power from a BS. ϕ_i^n is also a fraction of power assigned to $i-th$ MS among total transmission power from $0-th$ BS, or serving BS. The intercell interference of an MS which is located in normal region is derived as

$$I_{inter} = \sum_{k=1}^{K}\sum_{n=1}^{J}\alpha^n\cdot P_k\cdot L_k = \sum_{k=1}^{K}P_k\cdot L_k \tag{12}$$

The total transmission power of a BS is determined by the number of on-going calls served by the BS. In this paper, we assume the worst situation that all BS's serve maximum number of on-going calls and transmit signals at full power level. And, the interested MS is located at the cell boundary. Then, the E_b/N_t of the received signal at the MS is found as

$$(\frac{E_b}{N_t}) = \frac{W}{R} \cdot \sum_{j=1}^{J^{'}(0)} \frac{\alpha^j \cdot \psi \cdot \phi_i^n \cdot P_0 \cdot L_0}{I_{intra} + I_{inter} + N_0 \cdot W}$$
$$\approx \frac{W \cdot \psi}{R} \cdot \sum_{j=1}^{J^{'}(0)} \frac{\alpha^j \cdot \phi_i^n \cdot P_0 \cdot L_0}{I_{intra} + I_{inter}} \qquad (13)$$

In the above equation, we ignore $N_0 \cdot W$, because the background noise is negligible compared to the total transmission power. Also, $\psi \cdot \phi_i^n \ll 1$. Then, the relative power of $i - th$ MS is given by

$$\phi_i^n = \frac{(E_b/N_t) \cdot R}{W \cdot \psi} \cdot \frac{1}{P_0 \cdot L_0 \cdot \sum_{j=1}^{J'(0)} \frac{\alpha^j}{I_{intra}+I_{inter}}}$$
$$\approx \frac{(E_b/N_t) \cdot R}{W \cdot \psi} \cdot \frac{1}{\sum_{j=1}^{J'(0)} (\frac{\alpha^j}{1-\alpha^j + \sum_{k=1}^{K} \frac{L_k}{L}})} \qquad (14)$$

3.3 Averaging Factor

Equation (10) and (14) which we derived are the relative power of an MS in the worst situation for soft and hard handoff modes. For MS's in handoff region, the worst case means that the MS is located at the center of handoff region and serving BS and target BS transmit signals with the same power. Using (10) and (14), we can calculate the average relative power of an MS in handoff region. The relative power of an MS is proportional to $I_{inter}/P \cdot L$ as shown in (10). For the case that the users are uniformly distributed, the ratio of interference from the $j - th$ adjacent BS to the received power from serving BS is given by

$$\frac{\overline{I_{inter,j}}}{S_0} = \int_0^{\theta} \int_{r_1}^{r_2} x \cdot \frac{I_{inter,j}(x)}{S_0(x)} \cdot f_{r,\theta}(x,y) dxdy \qquad (15)$$
$$= \frac{I_{inter,j}(r_2)}{S_0(r_2)} \cdot \int_0^{\theta} \int_{r_1}^{r_2} x \cdot (\frac{r_2 \sqrt{R_j^2 + x^2 - 2R_j x \cos(|\theta_j - y|)}}{x\sqrt{R_j^2 + r_2^2 - 2R_j r_2 \cos(|\theta_j|)}})^{-l} \frac{2}{r_2^2 \theta} dxdy$$

where r_2 is the radius of a cell, θ is the angle of a sector, and for the unsectorized cell, $\theta = 2\pi$.

$I_{inter}/(P \cdot L)$ for the case that users are uniformly distributed in the cell is η times that of the worst case that all users are at the center of handoff region. Hence, the average power factor in handoff region η is given by

$$\eta = \sum_j (\int_0^{\theta} \int_{r_1}^{r_2} x \cdot (\frac{r_2 \sqrt{R_j^2 + x^2 - 2R_j x \cos(|\theta_j - y|)}}{x\sqrt{R_j^2 + r_2^2 - 2R_j r_2 \cos(|\theta_j|)}})^{-l} \frac{2}{r_2^2 \theta} dxdy) \qquad (16)$$

where R_j and θ_j are the coordination of adjacent $j - th$ BS. The average power factor η was introduced in [9], and we extended this into the case that an MS is located in handoff region.

4 Numerical Results

In this section, the numerical results for relative transmission power assigned to an MS from serving BS are calculated. We consider the system model which consists of 12 circular cells and the radius of cell is 100m. MS's are uniformly distributed in each cell. In this numerical example, we consider three types of data traffic whose data rates are 153.6kbps, 307.2kbps, and 614.4kbps, respectively. We assume that a call has infinite amount of data. The target SIR is assumed to be 12.5dB. The fraction of BS power assigned to traffic channels (ψ) is 0.8, standard deviation of lognormal shadow fading is 8 dB and bandwidth W is $5MHz$. We also assume that the power portions for each RAKE receiver finger are $(0.8, 0.1, 0.05)$ and the ratio of handoff region in a cell is 0.3.

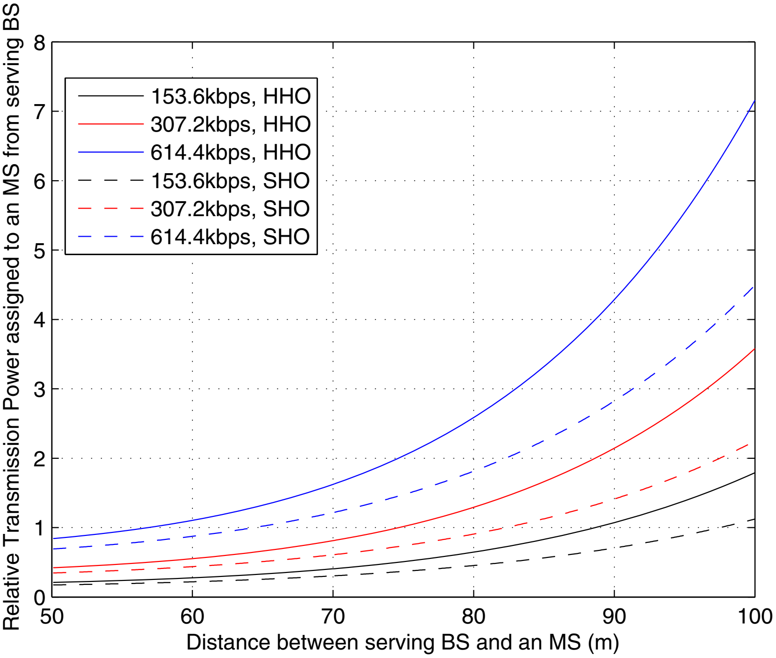


Fig. 3. Required relative transmission power of an MS

At first, we calculate the required relative transmission power of an MS according to its location. Fig. 3 shows the result. In this figure, the notation "HHO" indicates hard handoff mode and "SHO" means soft handoff mode. We can see that soft handoff requires lower transmission power than hard handoff for these three traffic. And the difference of required power increases as data rate of a traffic increases. The result indicates that soft handoff can reduce energy consumption of BS more than hard handoff.

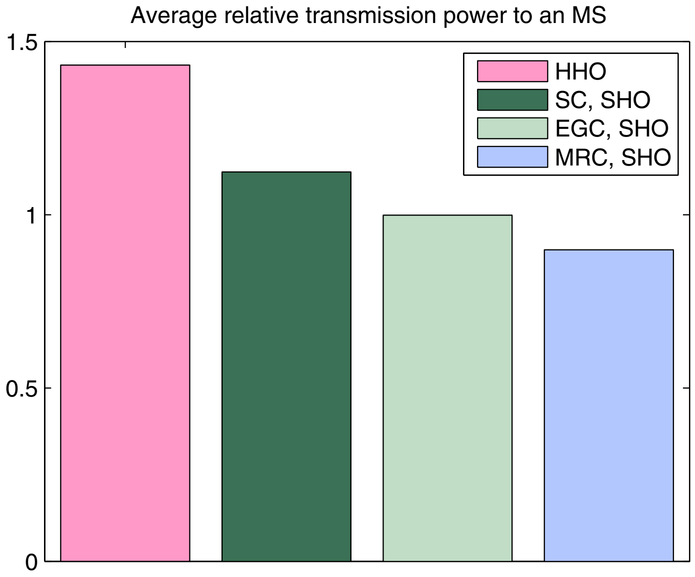


Fig. 4. Average relative transmission power according to combining method

We also investigate numerical results when various combining methods are used in acquiring diversity gain. Three combining method, such as selection combining (SC), equal gain combining (EGC), and maximal ratio combining (MRC) are considered. It is commonly known that the gain of each method is about $2.2dB$, $2.6dB$ and $3dB$, respectively [11]-[13]. The result is shown in Fig. 4.

Fig. 4 shows that the average relative power of MS's in handoff region increases as the value of combination gain decreases. Because the average relative power of MS's in handoff region should be increased to satisfy the target SIR value, as the value of combination gain decreases.

From these results, we can observe that soft handoff has better energy efficiency than hard handoff. As higher data rate is provided, the effect of power saving enlarges. This observation indicates that the use of soft handoff can help to save energy in RAN though it requires more code and radio resource.

5 Conclusions

In this paper we investigate the energy saving potential of handoff mode in mobile communication systems. The transmission power to an MS during handoff is analyzed for both hard handoff mode and soft handoff mode. As numerical results, we calculate the average relative transmission power of an MS who is provided voice, non-real-time data, and real-time multimedia traffic. We also compare numerical results when various combining methods are used in acquiring diversity gain. From these results, we can see that soft handoff requires lower transmission power than hard handoff. As higher data rate is provided, the effect of power saving between soft and hard handoff mode enlarges. This observation indicates that soft handoff mode is more efficient in view of energy consumption.

References

1. Green IT/Broadband and Cyber-Infrastructure, http://green-broadband.blogspot.com
2. Global Action Plan, An inefficient truth Global Action Plan Rep. (2007), http://www.globalactionplan.org.uk/
3. Node B datasheets (2008), http://www.motorola.com/
4. Louhi, J.T.: Energy efficiency of modern cellular base stations. In: INTELEC, Rome, Italy (2007)
5. Richter, F., Fehske, A.J., Fettweis, G.P.: Energy Efficient Aspects of Base Station Deployment Strategies for Cellular Networks. In: IEEE Vehicular Technology Conference VTC 2009, Anchorage, USA (200)
6. Brydon, A., Heath, M.: Will 3G Networks Cope?. Telecoms Market Research (2009)
7. Cheng, M., Chang, L.F.: Uplink system performance of high-speed IS-95 CDMA with mixed voice and bursty data traffic. In: Proc. IEEE PIMRC 1998, Boston, MA, pp. 1018–1022 (1998)
8. Chung, Y., Cho, D.H.: Introduction to a New Performance Factor of Soft Handoff for Real-time streaming services. IEICE Trans. Commun. E89-B(10), 2933–2935 (2006)
9. Jansen, M.G., Prasad, R.: Capacity, throughput, and delay analysis of a cellular DS CDMA system with imperfect power control and imperfect sectorization. IEEE Trans. Veh. Technol. 44(1), 67–75 (1995)
10. Choi, W., Kim, J.Y.: Forward-Link Capacity of a DS/CDMA System with Mixed Multirate Sources. IEEE Trans. Veh. Technol. 50(3), 737–749 (2001)
11. Lee, W.: Smaller Cells for Greater Performance. IEEE Commun. Mag. 29, 19–23 (1991)
12. Papen, W.: Improved Soft Handoff and Macro-Diversity for Mobile Radio. In: Proc. IEEE ICC 1995, Seattle, WA, pp. 1828–1833 (1995)
13. Gorricho, J.-L., Rojas, A., Paradells, J.: Power control at the combiner output to maximize the uplink capacity on a cellular spread spectrum system. IEEE Commun. Lett. 2(10), 273–275 (1998)

Energy Efficient Coexistence of WiFi and WiMAX Systems Sharing Frequency Band

Jongwoo Kim[1], Suwon Park[1], Seung Hyong Rhee[2], Yong-Hoon Choi[3], and HoYoung Hwang[4]

[1] Dept. of Electronics and Communications Engineering
[2] Dept. of Electronics Convergence Engineering
[3] Dept. of Information Control Engineering
[4] Dept. of Computer Engineering,
Kwangwoon University, Seoul, Korea
{jongwoo_kim,spark,rhee,yhchoi,hyhwang}@kw.ac.kr

Abstract. Various wireless communication systems in a shared frequency band such as 2.4GHz ISM band are operating. This causes the mutual interference among the wireless communication systems, and makes worse the performance of each of them. They should use more energy to achieve the desired quality of service. Many studies have been carried out to solve the mutual interference problem, called the coexistence problem. In this paper, we quantitatively analyze the effect of the mutual interference between Wi-Fi and WiMAX systems, and propose a method to solve the problem and evaluate its performance by simulation.

Keywords: Coexistence, mutual interference, power saving mode, PS-Request.

1 Introduction

To support various users' diverse services, various types of wireless communication systems are needed. Among the wireless communication systems, several wireless communication systems share a frequency band such as 2.4GHz industrial, scientific and medical (ISM) band. Well known wireless communication systems in the ISM band are WiFi as wireless local area network (WLAN) system, and Bluetooth and ZigBee as wireless personal area network (WPAN) systems. The coexistence in the shared frequency band causes mutual interference among the wireless communication systems, and makes worse the performance of each wireless communication system. Many researches have been carried out to reduce or avoid the mutual interference, so called coexistence problem. The coexistence of WiFi and Bluetooth systems is dealt in [1], [2] and [3], that of WiFi and ZigBee is done in [4], [5].

Recently, WiMAX subscribers at home or in office want to use high speed wireless internet service with low price or for nothing. As a result, WiMAX system is being considered as another candidate system operated in the ISM band. Because WiFi system is one of the popular wireless communication systems in the ISM band and is deployed in many indoor or outdoor places, WiFi and WiMAX systems sharing a frequency band may be operated in adjacent or the same area as shown Fig. 1. In that

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 164–170, 2010.

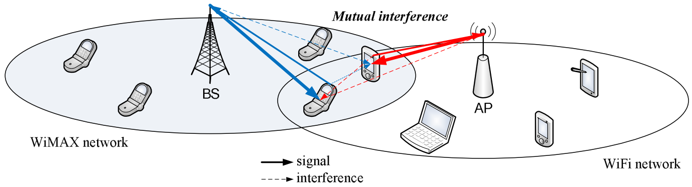


Fig. 1. Coexistence of WiFi and WiMAX system

case, they may interfere mutually. This may degrade the performance of each system such as frame error rate (FER), block error rate (BLER) or throughput. Consequently, many studies for coexistence of WiFi and WiMAX systems have been carried out to solve the coexistence problem. The studies are classified as two categories. One is how much interference is there between them [6][7], and the other is how to avoid or mitigate the mutual interference [8][9][10][11][12][13].

2 Proposed Solution

Fig. 2 shows a desiring time division operation (TDO) of WiFi and WiMAX systems.

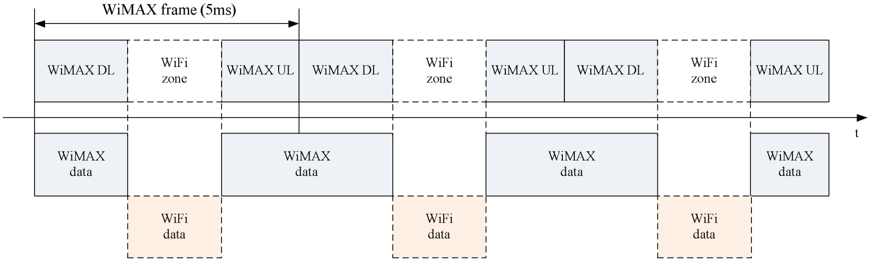


Fig. 2. Time Division Operation of WiFi and WiMAX Systems

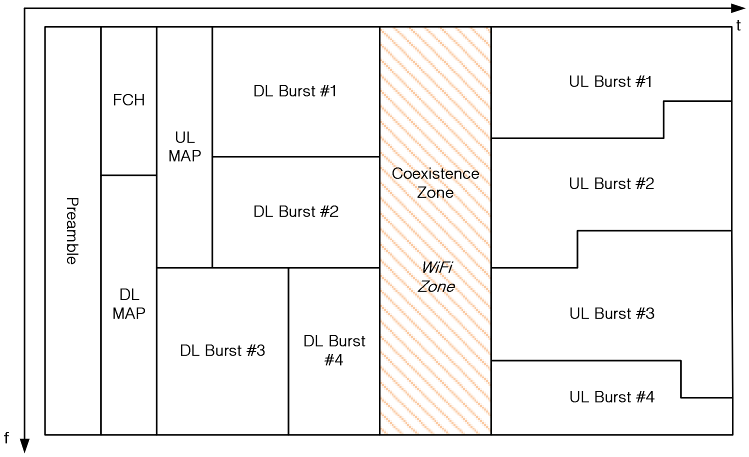


Fig. 3. WiFi transmission duration in WiMAX frame structure

For the TDO such as Fig. 2, the master system such as BS of WiMAX or AP of WiFi should control the transmission of its terminals. But conventional WiFi system can not control the transmission of STAs. Thus, we proposed a novel PS-Request protocol which makes the AP control the transmission of STAs [8]. It was based on use of vestigial power management bit within WiFi frame structure of access point (AP) which has not been used in the conventional power saving mode (PSM) of WiFi system, and coexistence zone such as WiFi zone shown in Fig. 3 was proposed for TDO of WiMAX and other systems. By using the proposed coexistence zone, the proposed PS-Request protocol guarantees the transmission duration of each system without the mutual interference. In this paper, the performance of the proposed one will be evaluated by computer simulation.

3 Simulation and Result

We allocate the transmission duration to each system as shown in Fig. 3. We assume that the transmission duration of WiFi system is 1.8432ms, and that of WiMAX system is 2.9952ms. The transmission duration is allocated based on the OFDM symbol duration of WiMAX system.

3.1 Simulation Conditions

Table 1 shows used parameters for computer simulation.

Table 1. Simulation parameters

	WiMAX	WiFi
Bandwidth(MHz)	8.75	20
Sampling Frequency	10	20
Over-sampling Ratio	2	1
Extended Sampling Frequency (MHz)	20	20
FFT Size	1024	64
Subcarrier Allocation	PUSC	-
Pulse Shaping Filter	Raised Cosine (Roll-off factor=0.25)	Raised Cosine (Roll-off factor=0.25)
Channel Coding	Convolutional coding (R=1/2, K=7)	Convolutionalcoding (R=1/2, K=7)
Modulation	QPSK	QPSK
Subcarrier Space (kHz)	9.765	312.5
Channel	AWGN	AWGN
Allocated Time Interval	2.9952 ms	1.8432 ms

3.2 Simulation Results

Higher SIR (Signal-to-Interference Ratio) in the figures means that at the interferee, average received signal power for the interferee is relatively larger than the average received interference power from the interferer. The relative quantity of interference is inversely proportional to the SIR shown in the legend of the figures. “without interference” in the legend is equivalent to SIR = ∞ dB.

Fig. 4 and 5 shows the performance of WiMAX and WiFi systems not using the proposed PS-Request protocol, respectively.

Fig. 4 shows the throughput vs. received E_b/N_0 of WiMAX system under the interference from WiFi system sharing the same frequency band. The noise spectral density N_0 in the received E_b/N_0 does not contain the interference for separately analyzing only the effect of the interference. For simplicity, it is assumed that WiMAX system consists of one BS and one MS, and all of the radio resources are allocated to the MS.

Fig. 5 shows the throughput vs. received E_b/N_0 of WiFi system under the interference from WiMAX system sharing the same frequency band [14].

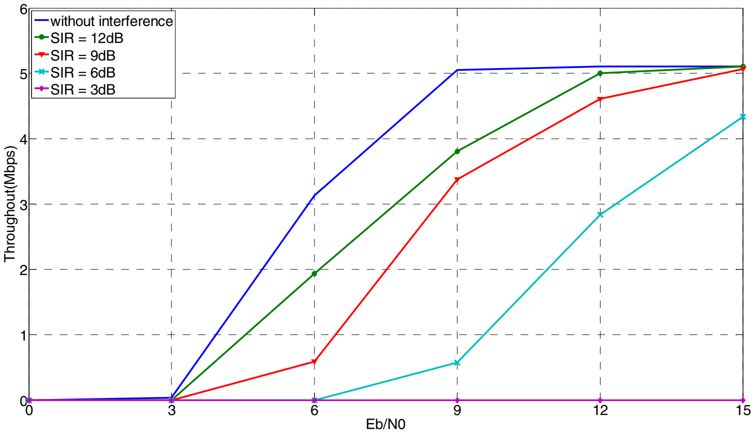


Fig. 4. WiMAX Throughput (Intereferee = WiMAX, Interferer = WiFi)

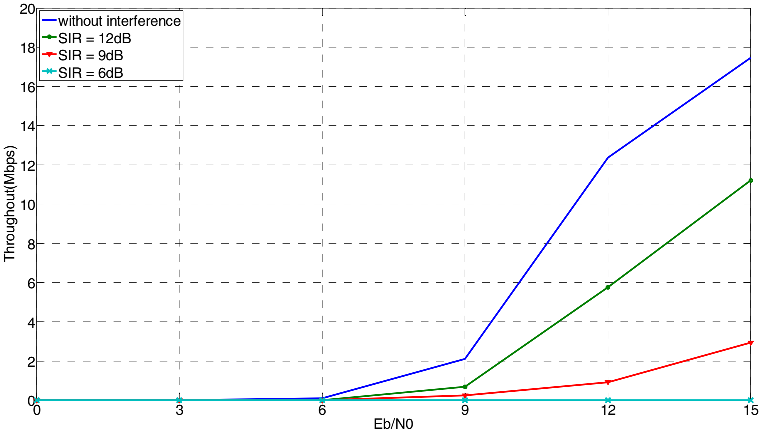


Fig. 5. WiFi Throughput (Intereferee = WiFi, Interferer = WiMAX)

Fig. 6 and 7 show the performance of WiMAX and WiFi systems using the proposed PS-Request protocol, respectively.

Fig. 6 shows the throughput vs. received E_b/N_0 of WiMAX system under the interference from WiFi system sharing the same frequency band.

Fig. 7 shows the throughput vs. received E_b/N_0 of WiFi system under the interference from WiMAX system sharing the same frequency.

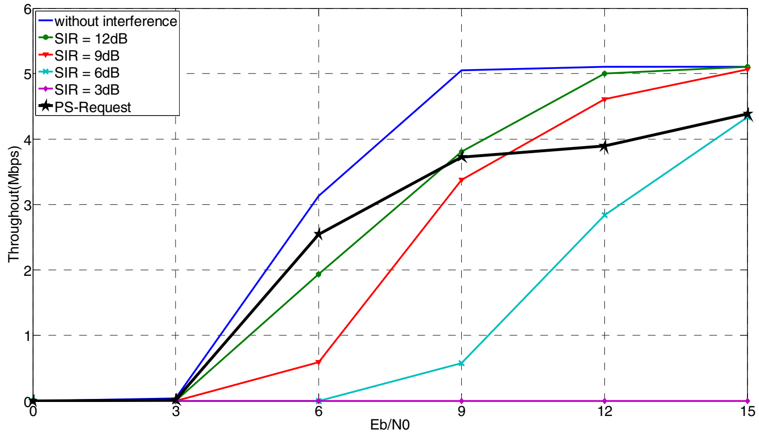


Fig. 6. WiMAX Throughput (Intereferee = WiMAX, Interferer = WiFi)

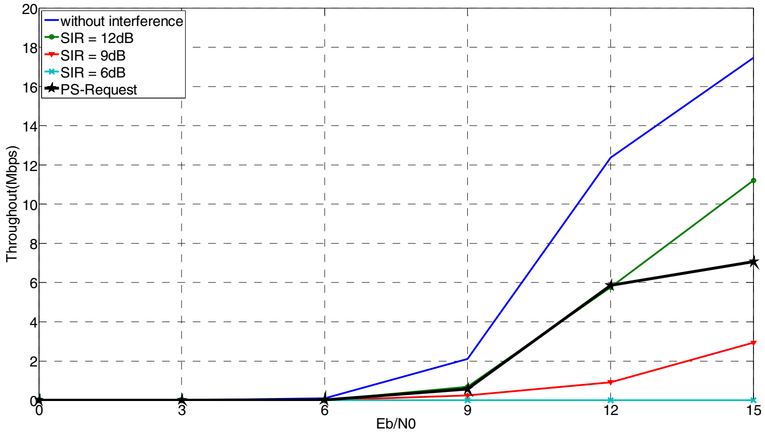


Fig. 7. WiFi Throughput (Intereferee = WiFi, Interferer = WiMAX)

The throughput of each system is decreased due to the mutual interference. Higher interference corresponding to smaller SIR causes smaller throughput for the same received Eb/No. As shown in Fig. 6 and 7, the proposed PS-Request based TDO has better performance in moderate or large level of interference environment. For small level of interference environment including interferenceless (without interference) environment, it can have worse performance because allowed transmission interval for each system can not be used by the other system.

4 Conclusion

In this paper, we evaluated the performance of the proposed PS-Request protocol using the vestigial power management bit of AP of WiFi system for the coexistence of WiFi and WiMAX systems in a shared frequency band by computer simulation. For smaller or zero interference environment, the proposed one can have worse performance such as smaller throughput because allowed transmission interval for each system can not be used by the other system. However, for smaller or zero interference environment, the proposed one can have better performance such as relatively higher throughput.

Acknowledgments. This work was partly supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2009-0086288), and the present research has been partly conducted by the Research Grant of Kwangwoon University in 2009.

References

1. Chiasserini, C.F., Rao, R.R.: Coexistence Mechanisms for Interference Mitigation between IEEE 802.11 WLANs and Bluetooth. In: Proceedings of INFOCOM 2002, pp. 590–598 (2002)
2. Glomie, N., Chevrollier, N., Rebala, O.: Bluetooth And Wlan Coexistence: Challenges And Solutions. IEEE Trans. Wireless Commun. 10, 22–29
3. Hsu, A.C.-C., Wei, D.S.L., Jay Kuo, C.-C.: Coexistence Mechanism Using Dynamic Fragmentation for Interference Mitigation between Wi-Fi and Bluetooth. In: MICOM 2006 (2006)
4. Sikora, A., Groza, V.F.: Coexistence of IEEE 802.15.4 with other Systems in the 2.4GHz-ISM-Band. In: Proceedings of IEEE Instrumentation & Measurement Technology Conference, pp. 1786–1791 (2005)
5. Yuan, W., Wang, X., Linnartz, J.-P.M.G.: A Coexistence Model of IEEE 802.15.4 and IEEE 802.11b/g. In: Philips research (2007)
6. Kim, J., Kim, D., Park, S., Rhee, S.H.: Interference Analysis of Wi-Fi System and WiMAX System in Shared Band: Interference from WiMAX System to Wi-Fi System. In: Proceedings of The 19 Joint Conference of Communication and Information (2009)
7. Kim, D., Kim, J., Park, S., Rhee, S.H., Kang, C., Han, K., Kang, H.: Interference Analysis of Wi-Fi System and WiMAX System in Shared Band: Interference form Wi-Fi System to WiMAX System. In: Proceedings of The 19 Joint Conference of Communication and Information (2009)
8. Kim, J., Kim, D., Park, S., Rhee, S.H., Han, K., Kang, H.: Use of Vestigial Power Management Bit within Wi-Fi Frame Structure of Access Point for Coexistence of Wi-Fi and WiMAX Systems in Shared Bands. In: Proceedings of The First International Conference on Ubiquitous and Future Networks, pp. 220–224 (2009)

9. Kim, D., Kim, J., Park, S., Rhee, S.H., Kang, C., Han, K., Kang, H.: Circulator-Based Collocated System for Coexistence of Wi-Fi and WiMAX Systems in Shared Bands. In: Proceedings of The First International Conference on Ubiquitous and Future Networks, pp. 214–219 (2009)
10. Berlerman, L., Hoymann, C., Hiertz, G.R., Mangold, S.: Coexistence and Interworking of IEEE 802.16 and IEEE 802.11(e). In: IEEE 63rd Vehicular Technology Conference, VTC 2006, vol. 1, pp. 27–31 (Spring 2006)
11. Berlerman, L., Hoymann, C., Hiertz, G.R., Walke, B.: Unlicensed Operation of IEEE 802.16: Coexistence With 802.11(a) in Shared Frequency Bands. In: IEEE 17th International Symposium Personal Indoor and Mobile Radio Communications, pp. 1–5 (2006)
12. Jing, X., Raychaudhuri, D.: Spectrum Co-existence of IEEE 802.11b and 802.16a Networks Using Reactive and Proactive Etiquette Policies. In: 2005 First IEEE International Symposium New Frontiers in Dynamic Spectrum Access Networks, DySPAN 2005, pp. 243–250 (2005)
13. Jing, X., Mau, S.-C., Raychaulhuri, D., Matyas, R.: Reactive Cognitive Radio Algorithms for Co-existence between IEEE 802.11b and 802.16a Networks. In: IEEE Global Telecommunications Conference, GLOBECOM 2005, vol. 5, pp. 2465–2469 (2005)
14. Kim, J., Park, S., Choi, Y.-H., Rhee, S.H.: Performance of Wi-Fi System due to Interference from WiMAX System in a Shared Frequency Band. In: Proceedings of The China-Korea Joint Conference on Information and Communications, JCIC 2010, pp. 103–104 (2010)

IP Mobility Performance Enhancement Using Link-Layer Prediction

Jun-Hui Lee, Hyun-Woo Kim, Yong-Hoon Choi[⋆,⋆⋆],
Young-Uk Chung, and Seung-Hyong Rhee

School of Electronics and Information Engineering,
Kwangwoon University,
447-1, Wolgye-dong, Nowon-gu,
Seoul 139-701, Korea
{yhchoi,yuchung,rhee}@kw.ac.kr

Abstract. In this paper, a prediction-based *L2 Trigger* approach is proposed for enhancing the performance of IP mobility in an integrated mobile Internet (e.g., Mobile WiMAX) and fast mobile IPv6 (FMIPv6) environment. The time series model of auto-regressive integrated moving average (ARIMA) is used to make short-term forecasting of mobile user's signal strength. Through the forecast of the signal strength, layer-3 handover activities occur prior to the start of layer-2 handover process, and therefore, total handover latency as well as service disruption time can be reduced.

Keywords: FMIPv6, prediction, cross-layer, time series analysis.

1 Introduction

One of the key challenges both in 3G and Mobile WiMAX is to provide seamless service for users moving at vehicular speeds. In order to support IP mobility, Mobile IP (MIP) has been adopted by 3GPP and WiMAX Forum. However, due to the very long handover latency of MIP, real-time services such as video streaming and VoIP are still hard to provide on those mobile networks.

Recently, cross-layer design techniques have been widely adopted for the purpose of reducing built-in delay of MIP. The protocols [1] and [2] enable a mobile station (MS) to quickly detect its movement into a new subnet by providing the new access router (AR) and the associated subnet prefix information when the MS is still connected to its current subnet. They commonly require *L2 trigger* as an early notice of an upcoming change in the layer-2 point of attachment. *L2 Trigger* can be utilized by the MS to start layer-3 handover-related activities in parallel with or prior to those of layer-2 handover. In work [3], the interaction

[⋆] Corresponding author.

[⋆⋆] This work was supported by the IT R&D program of MKE/KEIT. [KI002084, A Study on Mobile Communication System for Next-Generation Vehicles with Internal Antenna Array].

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 171–179, 2010.

between IEEE802.16 and Fast MIPv6 (FMIPv6) is presented with the primitives proposed by IEEE 802.21 for cross-layer handover design.

In this paper, we focus on issuing the appropriate *L2 Trigger* based on ARIMA prediction model. Using the signal strength samples obtained through scanning or periodic measurement report process, signal strength from serving base station (BS) and neighbor BS's can be predicted. Signal strength prediction is achieved by ARIMA(p, d, q) model without any assumption on the statistical properties of the movement. Although *L2 trigger* may come explicitly from MAC handover messages, it is more efficient to be derived from scanning process through the link layer prediction in terms of reducing handover latency and packet drops.

The rest of this paper is organized as follows. We first briefly describe the MAC-layer handover procedures specified in the IEEE standard, and the recent works on cross-layer handover protocols studied in the IETF standard. We then present our fast handover method based on ARIMA model together with experiment results. Finally, conclusions follow.

2 Background and Related Works

2.1 Layer 2: IEEE 802.16 Handover

The Mobile WiMAX system basically supports the *hard handover* (also known *as break-before-make*) scheme, but it also optionally supports the soft handover schemes such as macro-diversity handover (MDHO), and fast base station switching (FBSS). Handover is performed in two main processes: one is the network topology acquisition process and the other is handover execution process.

Network topology acquisition refers to periodically updating the parameter values needed for making handover decisions between the MS and the base station (BS). An MS may acquire neighbor BS information from a broadcast `MOB_NBR-ADV` message, or may actively scan target neighbor BS's and optionally try association in order to determine their suitability, along with other performance considerations as a handover target. The MS may incorporate information acquired from a `MOB_NBR-ADV` message to give insight into available neighbor BS's for cell reselection consideration.

A handover execution begins with a decision for an MS to handover from a serving BS to a target BS. The decision may originate either at the MS, or at the serving BS. The handover decision is notified to the BS through `MOB_MSHO-REQ` message or to the MS through `MOB_BSHO-RSP` message. The MS synchronizes to the DL transmissions of the target BS and obtain DL and UL transmission parameters. The MS and target BS may also conduct initial ranging. The final step in handover process is sending `MOB_HO-IND` message to the serving BS.

MS conducts network re-entry process, which is identical to initial network entry process, right after sending `MOB_HO-IND` message. Network re-entry process, however, may be shortened by the target BS's possession of MS information obtained from the serving BS over the backbone network. Network re-entry process completes with re-establishment of provisioned connections.

2.2 Layer 3: Recent Works on Cross Layer Handover Design

Mobile IP Low Latency Extension. Reference [2] proposed three methods to achieve low-latency MIP handovers: *pre-registration*, *post-registration*, and *combined method*. The *pre-registration* approach allows the MS to communicate with the new foreign agent (nFA) while still connected to the old FA (oFA). Therefore, MS can pre-build its registration state on the nFA prior to an underlying layer-2 handover. The pre-building process is initiated by an *L2 trigger*, which is an early notice of an upcoming change in the L2 point of attachment of the mobile node to the access network. Standard MIP registration process is performed between nFA and home agent (HA). If the registration is successful then packets for the MS are tunneled from the HA to the nFA where the MS has moved to.

The *post-registration* handover method proposes extensions to the MIP protocol to allow the oFA and nFA to utilize L2 triggers to set up a bidirectional tunnel between oFA and nFA that allows the MS to continue using its oFA while on nFA's subnet. This enables a rapid establishment of service at the new point of attachment which minimizes the impact on real-time applications. The MS must eventually perform a formal MIP registration after layer-2 communication with the new FA is established. Until the MS performs registration, the FAs will setup and move bidirectional tunnels as required to give the MS continued connectivity.

The *combined method* involves running a pre-registration and a post-registration handover in parallel. Reference [4] evaluated three low-latency schemes and compared their performances in terms of disruption time for VoIP services.

Fast Mobile (FMIP) IPv4/v6. Reference [1] proposed fast handover techniques, which eliminates signaling traffic between MS and HA. The handover mechanism is as follows:

1. MS receives proxy router advertisement (PrRtAdv) messages from the previous access router (PAR) either a solicited or unsolicited manner.
2. With the information provided by in the PrRtAdv message, the MS formulates a prospective new care-of-address (NCoA) and send a fast binding update (FBU) message, when it is still present on the PAR's link.

The purpose of FBU is to authorize PAR to bind previous CoA (PCoA) to NCoA, so that arriving packets can be tunneled to the new location of the MS. Fast binding acknowledgement (FBack) message is sent by PAR in response to FBU message. Depending on whether an FBack is received or not on the previous link, there are two modes of operation.

– *Predictive mode* of operation: The MS receives FBack on the previous link. This means that packet tunneling would already be in progress by the time the MS handovers to NAR. The MS should send unsolicited neighbor advertisement (UNA) message immediately after attaching to NAR, so that arriving as well as buffered packets can be forwarded to the MS right away.

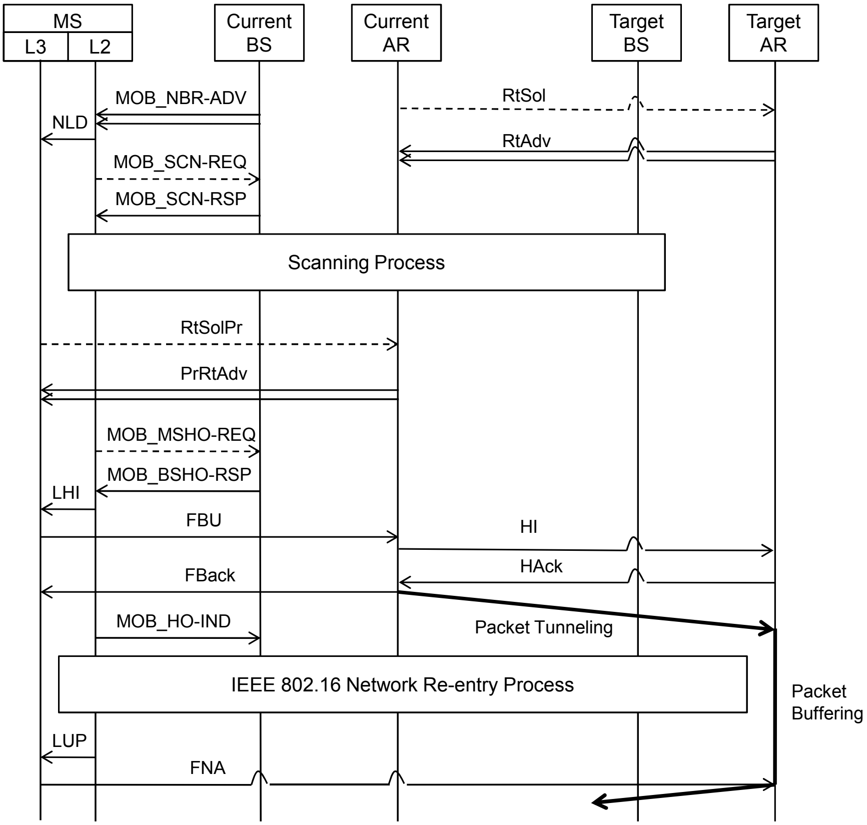


Fig. 1. Message sequence diagram of integrated FMIPv6 and IEEE 802.16 networks: *predictive-mode* operation scenario

- *Reactive mode* of operation: The MS does not receive FBack on the previous link. One reason for this is that the MS has not sent the FBU. The other is that the MS has left the link after sending the FBU, but before receiving an FBack. The MS announces its attachment immediately with an unsolicited neighbor advertisement (UNA) message that allows the NAR to forward packets to the MN right away, so that arriving as well as buffered packets can be forwarded to the MS right away.

Reference [3] describes FMIPv6 handovers on IEEE802.16e networks. In work [3], the interaction between IEEE802.16e and FMIPv6 is presented with the primitives proposed by IEEE 802.21 for cross-layer handover design. An example handover scenario utilizing the L2 triggers (e.g., New Link Detected (NLD), Link Handover Impend (LHI), and Link Up (LUP)) in FMIPv6 over Mobile WiMAX environment is illustrated in Fig. 1.

3 L2 Trigger Using Signal Strength Prediction

Let $\{x_m(t)|1 \leq t \leq n\}$ is the time series of measured received signal strength indiction (RSSI) values for which we want to predict their amount between a given MS and BS_m as shown in Fig. 2.

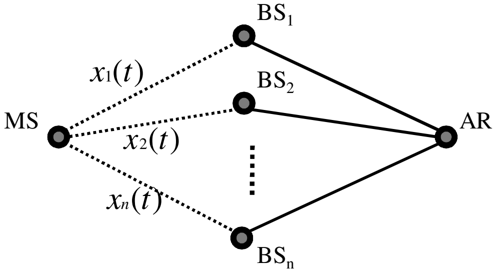


Fig. 2. Time series of signal strength $x_i(t)$

In order to obtain the time series, we monitor signal strength between a given MS and neighbor BS's at regular interval. The time series $\{x_m(t)|1 \leq t \leq n\}$ is expressed by previous observations $x_m(t-i)$, and noise term $e_m(t)$ which typically correspond to external events. The noise processes $e_m(t)$ are assumed to be uncorrelated with a zero mean and finite variance. The general ARIMA(p, d, q) model has the form

$$\phi(B)\nabla^d x_m(t) = \theta(B)e_m(t) \tag{1}$$

where B is the backward-shift operator defined by $B^j x_m(t) = x_m(t-j)$ and $B^j e_m(t) = e_m(t-j)$. $\nabla^d = (1-B)^d$ is the d^{th} order difference operator. $\phi(B)$ and $\theta(B)$ are the auto-regressive (AR) and moving average (MA) operators of order p and q, respectively, which are defined as $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q$, where $\phi_i(i = 1, 2, ..., p)$ are the AR coefficients, and $\theta_j(j = 1, 2, ..., q)$ are the MA coefficients.

According to the Box-Jenkins methodology [5], we identify the model parameters (p, d, and q) and coefficients (ϕ_i and θ_j). To make a prediction, minimum mean square error (MMSE) forecast method is used. Let's denote the k-step-ahead prediction as $\widehat{x}_m(t+k)$. In this paper, we performed one-step-ahead prediction. Each time the MS obtains a new RSSI sample, the coefficients $\phi_i(i = 1, 2, ..., p)$ and $\theta_j(j = 1, 2, ..., q)$ are updated and corresponding predicted RSSI value, $\widehat{x}_m(t+1)$ is obtained. If $\widehat{x}_s(t+1) \leq H_{Th}$ and $\widehat{x}_n(t+1) - \widehat{x}_s(t+1) \geq 3dB \quad for \quad \exists n$, *L2 Trigger* is issued. The subscripts s and n stand for serving BS and target BS, respectively. The threshold H_{Th} is the signal strength threshold to start layer-2 handover process. We assumed that hysteresis margin for preventing ping-pong is -3 dB.

4 Simulation Results

4.1 Mobility Models and Network Topologies

We assumed that each BS is connected to a different AR so that when an MS changes BS attachment, corresponding layer-3 handover is always required. We consider three mobility models - *Manhattan*, *Freeway*, and *Random Waypoint* model.

In *Manhattan Mobility Model (MM)* [6] the vehicle is allowed to move along the grid of horizontal and vertical streets. At an intersection the vehicle can turn left, right or go straight with probability 0.25, 0.25, and 0.5, respectively. The velocity of a vehicle at a time slot is dependent on its velocity at the previous time slot. In our simulation, we use the following parameters: total grid area is $5,000m \times 5,000m$, road-to-road distance is 50m, average speed is 40km/h, BS-to-BS distance is 1,000m, and simulation time is 1h. In *Freeway Mobility Model (FM)* [6], a vehicle follows a highway in constant direction at average speed of 90km/h. As was in Manhattan model, the velocity of a vehicle is temporally dependent upon its previous velocity. In *Random Waypoint Model (RWM)*, a vehicle moves along the grid of horizontal and vertical roads in $5,000m \times 5,000m$ area, average speed is 20km/h, maximum pause time is 1s, and simulation time is 1h.

4.2 Numerical Results

For each particular mobility model, the signal strength values of a given MS and neighbor BS's including serving BS are periodically measured and represented as time series $x_i(t)$. In our simulation, we chose both sampling period and prediction interval as 100ms. For all time series $x_i(t)$, we used the first 80 samples (8-second data) to do the time series model fitting. We observed that the time series $x_i(t)$ is nonstationary. We obtained the first-order differencing of actual time series $x_i(t)$ to generate new times series, $w_i(t) = x_i(t) - x_i(t-1)$. This new time series was found to be stationary in the case of most of the data sets, and therefore, we fixed $d = 1$ in our ARIMA model. After analyzing ACF and PACF plots of $x_i(t)$, we observed that ARIMA$(1, 1, 1)$ is the best fitting model. Fig. 3 shows the observed RSSI and predictions that result from the ARIMA$(1, 1, 1)$ model for $x_s(t)$ and $x_n(t)$ in *Freeway* model. For the clarity of presentation, only 1,000 ms of data are shown in Fig. 3. It is seen that ARIMA$(1, 1, 1)$ follows the time series quite narrowly.

In our simulation, the predictive-mode operation probabilities of the proposed scheme are compared to those of the existing FMIPv6 protocol. We assumed that the existing FMIPv6 protocol start layer-3 handover process based on measured RSSI values, while the proposed scheme start layer-3 handover process if the equation (5) and (6) are satisfied. Note that the predictive-mode operation probabilities have a strong impact on the performance of total handover latency.

We conduct the experiments on the following cases: MAC frame length = 5ms, neighbor scan duration = 100ms, interleaving interval = 100ms, prediction

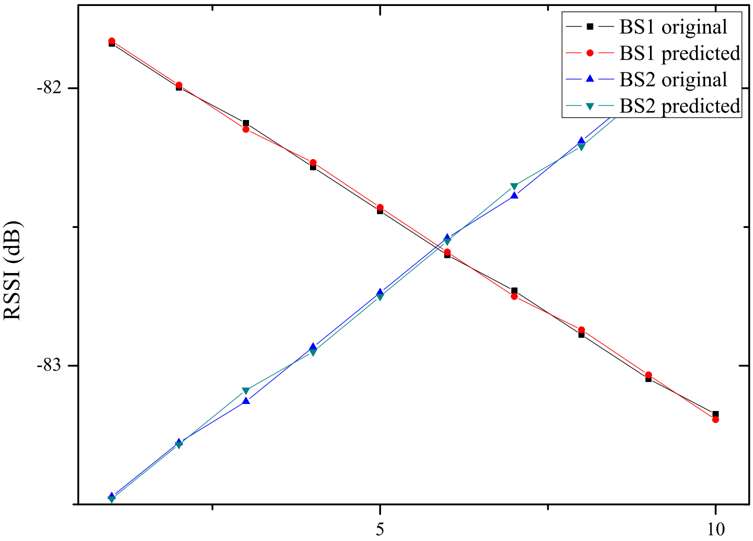


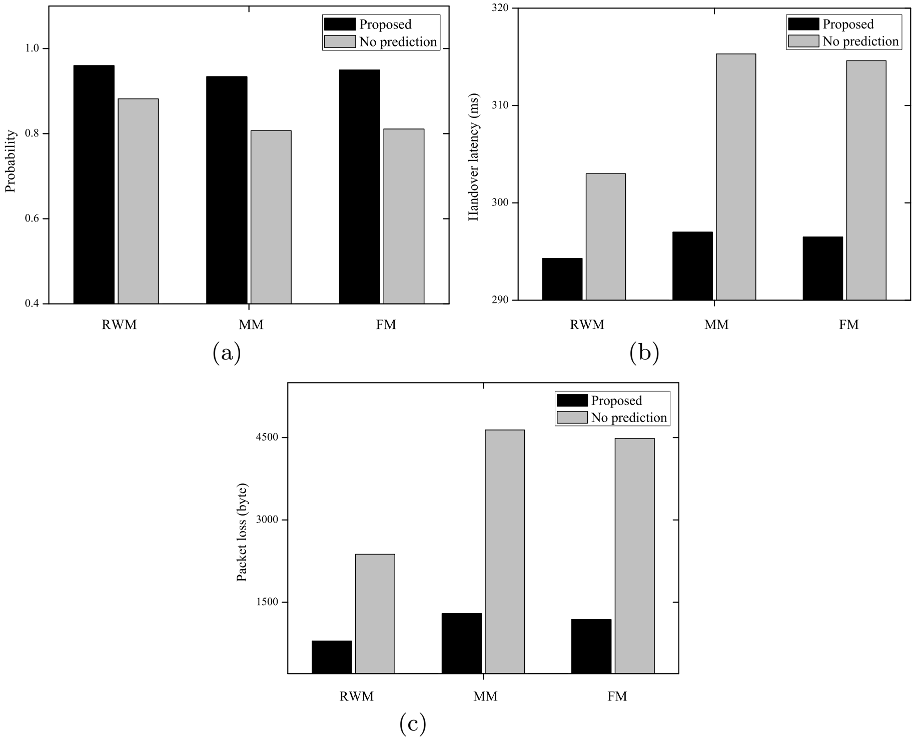
Fig. 3. The observed RSSI and predictions that result from the ARIMA(1,1,1) model for $x_s(t)$ and $x_n(t)$

interval = 100ms ahead, number of samples for calculating ϕ_i and θ_j: recent 80 samples, handover threshold H_{Th} = -80 dB, and handover hysteresis margin = -3dB. The results of the predictive-mode operation probabilities are given in Fig. 4(a).

When the randomness of the movement direction is low, causing more correct prediction, the improvement of predictive-mode probabilities is more visible in *Manhattan* and *Freeway* model. For example, the gain of *predictive-mode* operation probability in Freeway model is 0.95-0.81 = 0.14, while that of Random Waypoint model is just 0.96-0.88 = 0.08. Accordingly, the proposed scheme is more effective and more important in Freeway or Manhattan model than in Random Waypoint model.

Fig. 4(b) compares the handover latencies of two handover protocols with three different mobility models. The figure shows that original FMIPv6 has high handover latency due to low predictive-mode operation probabilities. Because the proposed scheme adopts forecasting method to increase *predictive-mode* operation probabilities, the handover latency decreases. For example, the handover latencies in our proposed scheme are below 297ms, while they are over 303ms in the original FMIPv6 for all three mobility models.

Fig. 4(c) compares the number of lost bytes during handover. For two handover schemes, there is no packet loss in predictive-mode because we assume that the buffer in NAR is not overflowed. The reactive mode still suffers packet drops because packets can be forwarded after the FBU message arrives at the PAR. Since the proposed scheme increases the predictive-mode operation probabilities, the total packet loss can be reduced.



(a)

(b)

(c)

Fig. 4. Comparison of (a) the probability of predictive-mode operation, (b) handover latency, and (c) packet loss

5 Conclusions

In this paper, we have presented ARIMA-based L2 Trigger approach in the integrated FMIPv6 and mobile Internet environment. Simulation results have shown that the proposed scheme outperforms existing conventional scheme in terms of handover latency and packet drops. Since the performance gains are more visible in MM and FM model, our approach is highly attractive for fast moving nodes.

References

1. Koodli, R.: Mobile IPv6 Fast Handovers. In: RFC 5568 (2009)
2. Malki, K.E.: Low-Latency Handoffs in Mobile IPv4. RFC 4881 (2007)
3. Jang, H.: et al.: Mobile IPv6 Fast Handovers over IEEE 802.16e Networks. RFC 5270 (2008)

4. Fathi, H., et al.: Mobility management for VoIP in 3G systems: evaluation of low-latency handoff schemes. IEEE Wireless Communications 12(2), 96–104 (2005)
5. Box, G., et al.: Time Series Analysis, Forecasting and Control, 3rd edn. Prentice-Hall, Englewood Cliffs (1994)
6. Bai, F., et al.: The Important framework for analyzing the Impact of Mobility on Performance of Routing protocols for Adhoc Networks. In: Ad Hoc Networks, vol. 1, pp. 383–403. Elsevier, Amsterdam (2003)

Header Compression for Resource and Energy Efficient IP over Tactical Data Link

Yongkoo Yoon[1], Suwon Park[1], Hyungkeun Lee[2],
Jong Sung Kim[3], and Seung Bae Jee[3]

[1] Dept. of Electronics and Communications Engineering, Kwangwoon University
[2] Dept. of Computer Engineering, Kwangwoon University
[3] Tactical Data Link PMO, Agency for Defense Development,
Seoul, Korea
ykooyoon@gmail.com, {spark,hklee}@kw.ac.kr,
jskim_add@yahoo.com, asherjee@gmail.com

Abstract. Link-16 is a tactical data link for providing information including tactical data link message necessary to a command and control unit. Even though current Link-16 transmits and receives only tactical data link message, there is a desire to send many kinds of IP applications over Link-16. However, too large size of IP header decreases the transmission efficiency of IP applications over Link-16. Thus, IP header compression algorithm suitable to Link-16 is needed without loss of performance. Conventional header compression algorithms such as RFC 2507, RFC 2508 and RFC 3095 are not suitable to Link-16 because it broadcasts information with low transmission data rate. In this paper, we propose a novel IP header compression algorithm for transmitting tactical data link messages and IP application data without modification of Link-16.

Keywords: Tactical Data Link, Link-16, IP, Header Compression.

1 Introduction

Future war concept is rapidly changing from the platform centric warfare to the network centric warfare. In the modern warfare, the quantity of owned platforms and weapons do not determine the victory of the war. Situation awareness (SA) and real-time command and control capability, and precise attack to the attack objective are more important. The network centric warfare uses all of these ones. In the network centric warfare, a surveillance system, command and control system, and attack system are networked, the situational awareness is shared among them, and joint and precise attack can be achieved. From them, the war performing ability can be maximized. [1]

The advantages of IP (Internet Protocol) over a tactical data link are as follows. First, services can be easily integrated. Integration of the tactical data link systems such as Link-16, Link-4A, Link-11, and the like, is easy. Second, new services can be easily and efficiently created by using open API (Application Programming Interface). Third, making broadband of the tactical communication network can be achieved. IP over

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 180–190, 2010.

tactical data link is a tactical communication network using latest communication technology. In the military communication network, the IP-based services can be popular like the commercial communication networks. Finally, various tactical data including multimedia tactical data can be easily introduced. The IP over the tactical data link can provide both raw tactical data and multimedia tactical data such as voice and image. From them, winning possibility in the war can be increased.

2 System Model Description

2.1 Network Model

Fig 1. shows a network model of IP over Link-16. Link-16 HOSTs are connected each other through Link-16 TDMA network. They use their own time slots. The role of DLP(Datalink Processor) in Figure 1 is a gateway between Link-16 TDMA network and IP network. It makes possible to communicate between IP servers in IP network and a Unit in Link-16 TDMA network.

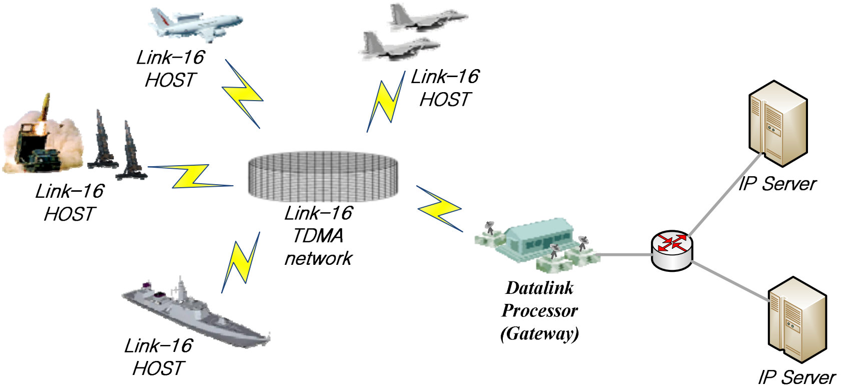


Fig. 1. Network Model of IP over Link-16

In this paper, the intra-network means a network including only Link-16 TDMA network, and the inter-network means a network including both Link-16 TDMA network and IP network connected by the gateway.

2.2 Data Transmission

Figure 2 describes a data transmission on IP over Link-16. A HOST (Unit 3 in Figure 2) in a Link-16 TDMA network doing a mission communicates the IP server located in either the internal Link-16 TDMA network or the external IP network connected by gateway. We call the former as intra-network case and the latter as inter-network case. In the TDMA network, a transport layer header, a network layer header and a Link-16 header are sequentially attached to the payload. [3] In the inter-network case, the data payload from the Link-16 unit passes through the gateway which transforms it to an IP packet and delivers it to the external IP network.

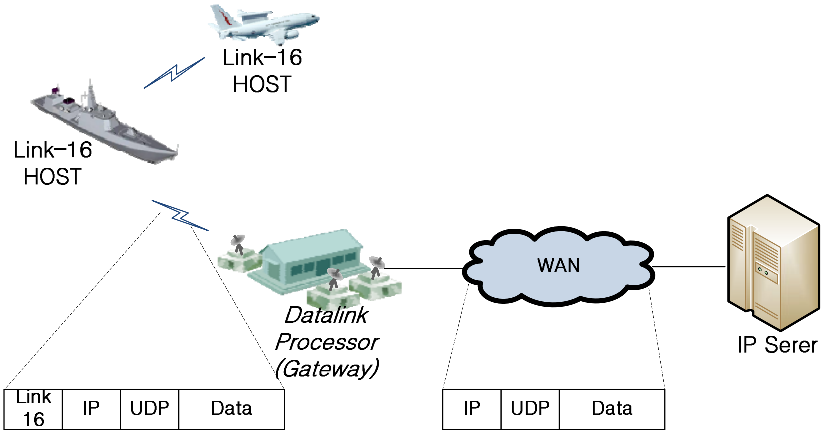


Fig. 2. UDP/IP Packet Header Structure for IP over Link-16

2.3 Header Overhead

Link-16 has four modes of data packing to a time slot; STDP (STandard Double Pulse), P2SP (Packed-2 Single Pulse), P2DP (Packed-2 Double Pulse), and P4SP (Packed-4 Single Pulse). Within a time slot, STDP, P2SP/P2DP and P4SP transmit 210, 420 and 840 bits, respectively. Transmitting data on IP over Link-16 is either a tactical data link message (Fixed-J series) or typical IP application message. The tactical data link message is composed of one to eight words, and each word is 70 bits. [4] As to the IP over Link-16, the problem that the transmission efficiency decreases is generated. Because the Link-16 in which the transmission capacity is little transmits the IP packet having a large size, transmission inefficiency is generated. For example, If Link-16 HOST transmits data by the packed-2 mode, 6 words(about 53 bytes) are transmitted at a timeslot. A UDP/IP header is 28bytes. Therefore it a header transmits by the packed-2 mode, it uses about 53% of the transmission capacity to a header and the rate of transmitting data is about 47%. This is a data transmission inefficiency. Thus UDP/IP header should be compressed for data transmission efficiency. That is the goal of this research.

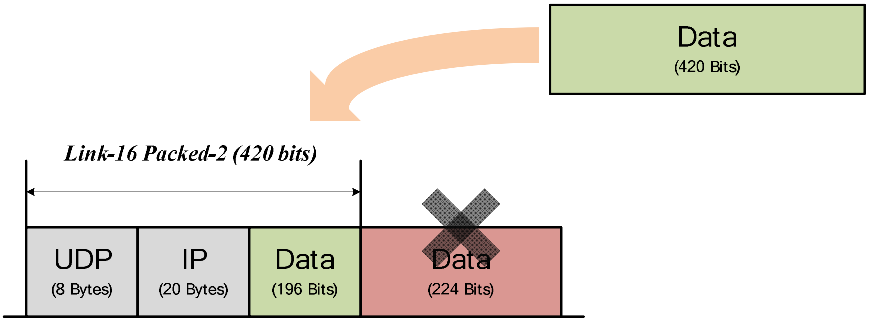


Fig. 3. Data Transmission Inefficiency by a UDP/IP Header

2.4 Header Compression Layer

Figure 4 describes a proposing compression layer for header compression suggested as a solution of the inefficiency of IP packet transmission over Link-16. Because the header size of the IP packet is too large, the IP packet moves to the compression layer, where it is compressed to the minimum size. The compression layer should reduce the size of the set of headers for IP packet such as UDP and IP headers, and output a compressed header. The receiver should recover the set of original headers such as UDP and IP headers based on the compressed header after a decompression procedure.

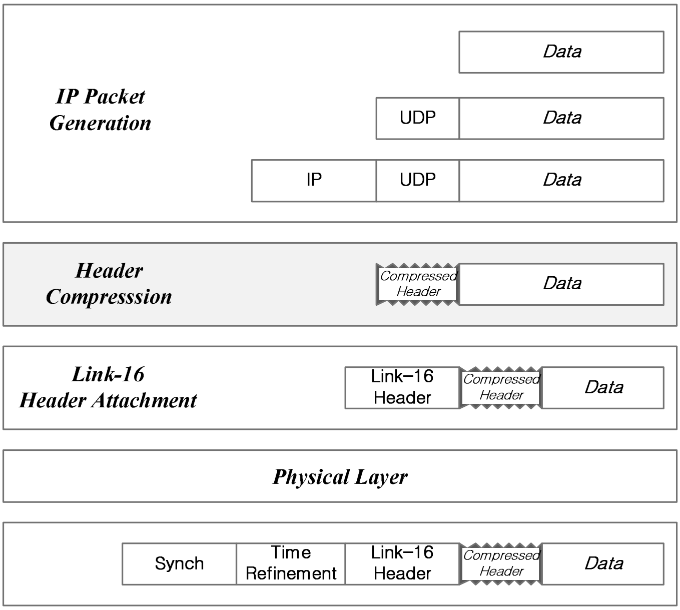


Fig. 4. Link-16 Layer Architecture for IP Packet Transmission

2.5 Existing Research

Header compression is not a new issue. In order to reduce the overhead of header, it has been studied for a long time in the field of wired or wireless communications, especially, for IP applications. RFC 2507, RFC 2508 are popular header compression algorithms. [5] They use delta coding which transmits the difference of the previous header and current header. [6][7] Even though the compression efficiency is high, they are weak for packet error that is common for all kinds of delta coding. That is, if a packet is erroneous, then the consecutive packets are not reliable.[8] But ROHC(Robust Header Compression : RFC 3095) is strong and data compression ratio is good for the packet error state. As to this, there is the feature which is strong against an packet error due to the LSB(Least Significant Bit) encoding not being delta coding.[9] Recently, it was studied as ROHCv2(RFC 4995, 4996, 5225) and it was applied to LTE(Long Term Evolution), that is 4G Mobile Communication candidate technology.[9]

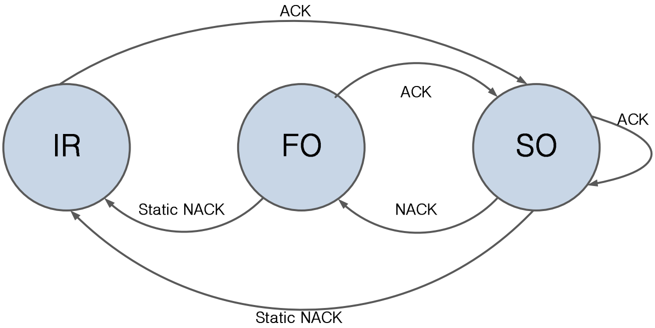


Fig. 5. Finite State Machine (FSM) of ROHC Compressor

ROHC which has the best compression efficiency among them is used in commercial mobile communications. Figure 5 shows the FSM(Finite State Machine) between compressor and decompressor. By using the feedback information from the compressor to the decompressor, the error of the compressed packet is checked. This operation is comprised based on the feedback channel. IR(Initializing and Refresh), FO(First Order), SO(Second Order) indicate the form of the compressed packet which a compressor transmits. IR is the full header that it transmits when a compressor starts a connection. SO is the state which compressor transmits the compressed packet after IR is transmitted without an error. FO is the compressed packet which it transmits in case the error of the SO packet was detected.

The robustness about packet error and the high data compression ratio of ROHC are generated from the adaptive change of operation mode. ROHC, there are three operation modes. First, U-mode(Unidirectional mode). The U-mode is the method being designed in order to operate in the unidirectional channel without the feedback channel. Second, O-mode(Optimistic mode) operates in the bidirectional channel. Compressor transmits the SO packet without ACK. Then, if the packet error is detected in the SO state, by using NACK, FSM is changed to FO or IR. Finally, it is R-mode(Reliable mode). In the SO state, R-mode receives ACK and NACK of all compressed packets then controls FSM. As to compressor, in O-mode, the transfer efficiency is high and the reliability is high in R-mode.

3 Proposing Header Compression Algorithm

3.1 Compressed Header Format

Two kinds of message can be transmitted on IP over a tactical data link such as Link-16. One is a tactical data link message, and the other is a typical IP packet. Link-16 message is a fixed format J-series message consisting of one to eight words. Each word is 210 bits. The IP packet payload is one of audio, image, text, tactical data, and the like.

After the compression layer processing shown in Figure 4 the compressed header for the tactical data link message is given in Figure 6.

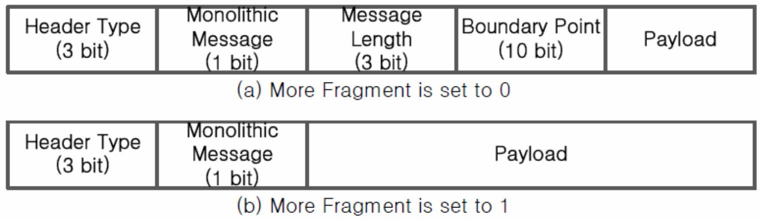

Header Type (3 bit)
Monolithic Message (1 bit)
Message Length (3 bit)
Boundary Point (10 bit)
Payload
(a) More Fragment is set to 0
Header Type (3 bit)
Monolithic Message (1 bit)
Payload
(b) More Fragment is set to 1

Fig. 6. Compressed Header Format for Tactical Data Link Message

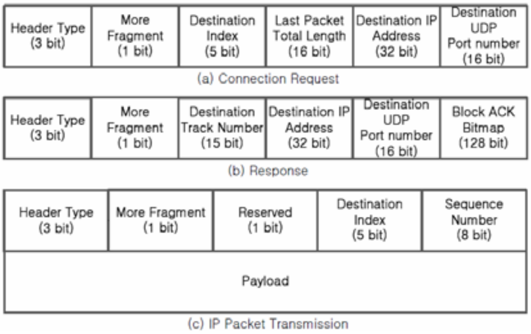
In the intra-network case, compressed header for IP over Link-16 is shown in Figure 6.

The header type field indicates the kind of a compressed packet. It is a 3-bit information and is explained in Table 1. The monolithic message field indicates whether one message is transmitted within a time slot or not. The message length field indicates the number of words for generated J-series message. It is used for recovering the total length value of IP header. The boundary point field indicates the boundary position when two messages are transmitted within one time slot. It prevents the consecutive error of a packet.

An IP packet over a tactical data link is delivered by using a point-to-point communication. Compressed header format for IP packet is described in Figure 7. It is composed of the connection request, response, and IP packet data transmission. A unit willing to transmit an IP packet sends the connection request packet which contains the information including IP address and port number, and an index number of the destination unit. The destination index is 5-bit information which is used for a receiver to identify its own packet. After a receiver receives the connection request packet, it sets transmitter information and transmits the response packet. If the response packet is received by using the assigned index number, the transmitter transmits a packet.

Table 1. Function classifications of the Header Type field

Header Type	Description
000	Tactical data link message transmission
001	IP packet transmission(ACK)
010	IP packet transmission(NACK)
011	Client connection request
100	ACK response
101	Server IP packet transmission(ACK)
110	Server IP packet transmission(NACK)
111	Reserved



Header Type (3 bit)	More Fragment (1 bit)	Destination Index (5 bit)	Last Packet Total Length (16 bit)	Destination IP Address (32 bit)	Destination UDP Port number (16 bit)

(a) Connection Request

Header Type (3 bit)	More Fragment (1 bit)	Destination Track Number (15 bit)	Destination IP Address (32 bit)	Destination UDP Port number (16 bit)	Block ACK Bitmap (128 bit)

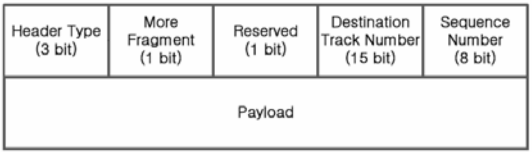
(b) Response

Header Type (3 bit)	More Fragment (1 bit)	Reserved (1 bit)	Destination Index (5 bit)	Sequence Number (8 bit)
Payload				

(c) IP Packet Transmission

Fig. 7. Compressed Header Format for IP Packet

The proposed header compression algorithm can select the use of ACK as shown in Table 1. It can complement the disadvantage of UDP protocol.



Header Type (3 bit)	More Fragment (1 bit)	Reserved (1 bit)	Destination Track Number (15 bit)	Sequence Number (8 bit)
Payload				

Fig. 8. Compressed Header Format for Data transmission of a Gateway

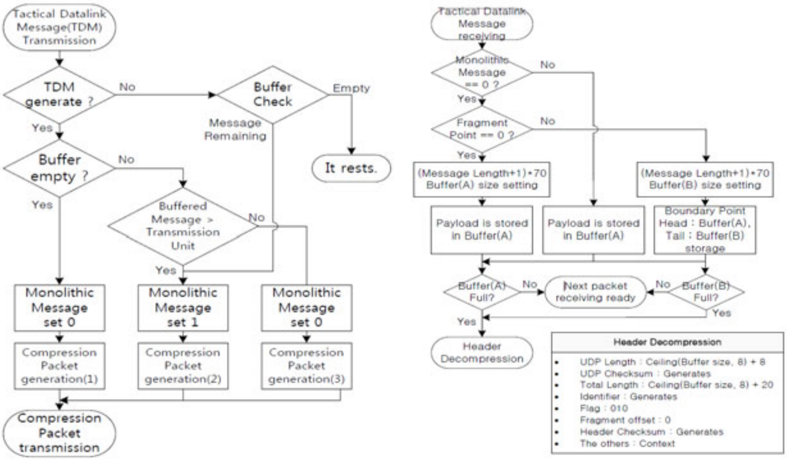
Transmission data format from an IP server is shown in Figure 8. The server uses the connection request and response packet, too. Instead of the index number, the server uses TN (Track Number) for its data transmission. The 5-bit index number can identify 32 connections, but 15-bit TN can identify 32768 connections. In general, a client connects to small number of servers, but a server connects to large number of clients. Thus, in the proposed scheme a client identifies 32 servers by using the index number, and a server identifies 32768 clients by using the TN.

3.2 Header Decompression

The objective of compression is to transmit reduced information and to restore the original from it. Within UDP and IP headers, the fixed valued field is removed and the changing valued field is used for compression. After decompression, all of the fields should be restored. Most of compressed protocols classify their fields into either fixed valued field or changing valued field. [6][7][8] Version, Header Length, Service Type, TTL, and Protocol fields are classified as fixed valued fields. The compression block of a transmitter does not transmit them. They are restored at the decompression block in the corresponding receiver. In the Link-16 network it is assumed that the constitutional unit knows IP Addresses corresponding to TN's. Thus the Source Address need not be transmitted. The Destination Address field is compressed during the connection request process and transmitted. The checksum field substitutes 12-bit CRC of Link-16. The Length field and Fragmentation field are remained. For the tactical data link message transmission, the Length field is compressed to the Message Length field within the compressed header.

3.3 Example of an Operation

The procedure of tactical data link message transmission using the proposed header compression algorithm is shown in Figure 9(a).


Tactical Datalink Message(TDM) Transmission
TDM generate ?
No
Yes
Buffer Check
Empty
Message Remaining
It rests.
Buffer empty ?
No
Yes
Buffered Message > Transmission Unit
No
Yes
Monolithic Message set 0
Monolithic Message set 1
Monolithic Message set 0
Compression Packet generation(1)
Compression Packet generation(2)
Compression Packet generation(3)
Compression Packet transmission
Tactical Datalink Message receiving
Monolithic Message == 0 ?
No
Yes
Fragment Point == 0 ?
No
Yes
(Message Length+1)•70 Buffer(A) size setting
(Message Length+1)•70 Buffer(B) size setting
Payload is stored in Buffer(A)
Payload is stored in Buffer(A)
Boundary Point Head : Buffer(A), Tail : Buffer(B) storage
Buffer(A) Full?
No
Next packet receiving ready
No
Buffer(B) Full?
Yes
Yes
Header Decompression
Header Decompression
UDP Length : Ceiling(Buffer size, 8) + 8
UDP Checksum : Generates
Total Length : Ceiling(Buffer size, 8) + 20
Identifier : Generates
Flag : 010
Fragment offset : 0
Header Checksum : Generates
The others : Context

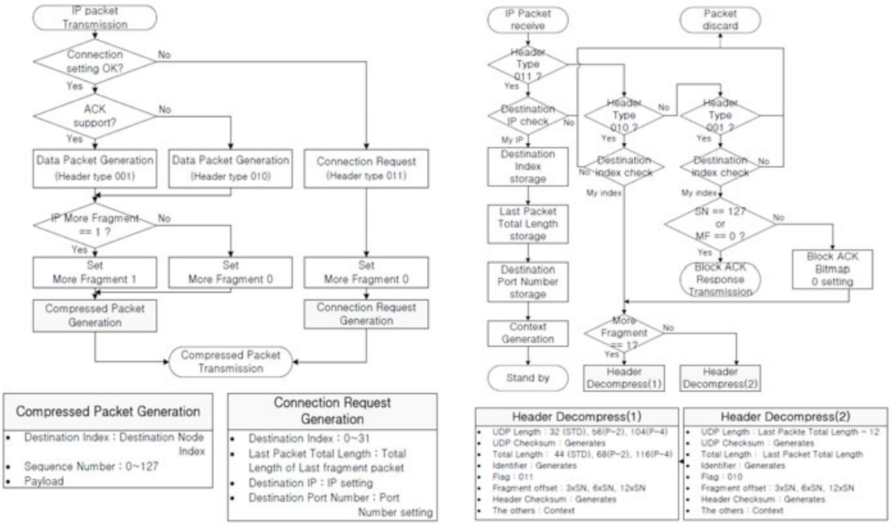
(a) Transmission (b) Reception

Fig. 9. Tactical Data Link Message (a) Transmission (b) Reception

A tactical data link message on IP over Link-16 is classified into one of three types of compressed packets.

- When there is no message in a buffer, the compressed packet (1) is a packet for new message.
- The compressed packet (2) is a packet with a monolithic message set to 1 and for enhancing the transmission efficiency.
- The compressed packet (3) is a packet containing both a part of message remained in the buffer and a new message.

Figure 9(b) describes the tactical data link message reception. According to the Monolithic Message field value and Boundary Point field value, the received message is classified into one of three types of packets. By using a context and a received compressed header, the original header can be restored. The context is a value known in advance. The checksum of transport and network layer, and Identifier value is restored at the receiver. Because they are not fragmented, the tactical data link message sets the Fragment offset value with 0. The Length field is set by the unit of byte.


IP packet Transmission
Connection setting OK?
No
Yes
ACK support?
No
Yes
Data Packet Generation (Header type 001)
Data Packet Generation (Header type 010)
Connection Request (Header type 011)
IP More Fragment == 1 ?
No
Yes
Set More Fragment 1
Set More Fragment 0
Set More Fragment 0
Compressed Packet Generation
Connection Request Generation
Compressed Packet Transmission
Compressed Packet Generation
Destination Index : Destination Node Index
Sequence Number : 0~127
Payload
Connection Request Generation
Destination Index : 0~31
Last Packet Total Length : Total Length of Last fragment packet
Destination IP : IP setting
Destination Port Number : Port Number setting
IP Packet receive
Packet discard
Header Type 011 ?
Yes
Destination IP check
No
My IP
Destination Index storage
Last Packet Total Length storage
Destination Port Number storage
Context Generation
Stand by
Header Type 010 ?
No
Yes
Destination index check
No
My index
Header Type 001 ?
Yes
Destination index check
No
My index
SN == 127 or MF == 0 ?
No
Yes
Block ACK Response Transmission
Block ACK Bitmap 0 setting
More Fragment == 1?
No
Yes
Header Decompress(1)
Header Decompress(2)
Header Decompress(1)
UDP Length : 32 (STD), 56(P-2), 104(P-4)
UDP Checksum : Generates
Total Length : 44 (STD), 68(P-2), 116(P-4)
Identifier : Generates
Flag : 011
Fragment offset : 3xSN, 6xSN, 12xSN
Header Checksum : Generates
The others : Context
Header Decompress(2)
UDP Length : Last Packte Total Length - 12
UDP Checksum : Generates
Total Length : Last Packet Total Length
Identifier : Generates
Flag : 010
Fragment offset : 3xSN, 6xSN, 12xSN
Header Checksum : Generates
The others : Context

(a) Transmission

(b) Reception

Fig. 10. IP Packet (a) Transmission (b) Reception

Figure 10(a) describes the IP packet transmission. For the IP packet transmission, a destination unit should be identified and a connection should be checked whether it is set or not. In case of no configuration, the connection request packet is transmitted. After the establishment of a connection, fragmented packets are sequentially transmitted, where the More Fragment bit of the last packet is set to 0.

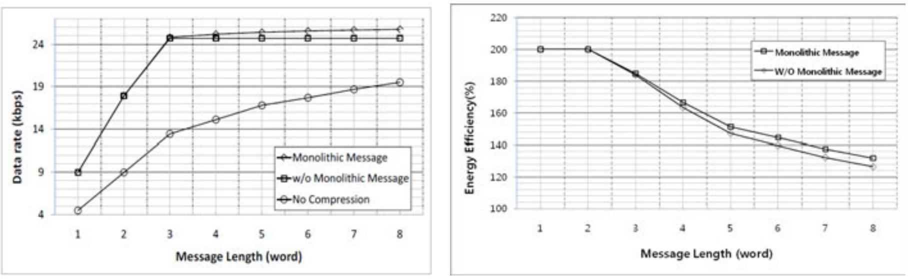
Figure 10(b) describes the IP packet reception. The IP packet is received based on a point-to-point communication. Therefore, it is important that deciding whether the received packet is its own or not. If it is not its own, it is discarded. According to the connection request and ACK request, it is classified as 3. The connection request packet saves the received information as a context which is used for header decompression. As to the ACK response packet, if the 128-th or the last packet is received, the received packet is transmitted. Feedback information is not transmitted if the ACK response mode is not used.

4 Simulation

The maximum data rate of Link-16 using the proposed header compression algorithm is evaluated.

Figure 11(a) shows the transmission data rate when the tactical data link message of one to eight words is transmitted. In case that the compressed header is not used, the UDP and IP headers are transmitted in a time slot. If a message containing three words is transmitted, the UDP and IP headers are transmitted to two time slot intervals. This reduces the maximum data rate (26.88 kbps) of STDP mode as a half. The difference of the data rate in case that the Monolithic Message field is used is shown. As the Monolithic Message transmits message containing more words, the relative size of a header is decreased. The data rate is larger because the message size is relatively larger.

Figure 11(b) shows the Energy efficiency according to Monolithic message field of header compression algorithm. STDP of link-16 transmits with 3 word units. Therefore, it has the transfer efficiency of two times in comparison with uncompressed mode until 2 words are transmitted. It is the average transmission efficiency 164.57% of Monolithic message and w/o Monolithic message is 161.561%. The difference of about 3% appears.



(a) Transmission Data Rate

(b) Energy Efficiency

Fig. 11. Performance for STDP (Fixed format J-series format) (a) Transmission Data Rate (b) Energy efficiency

Table 2 shows the IP packet transmission rate for two addressing methods to a unit. One is the IP addressing, and the other is the indexing scheme proposed in this paper. The IP addressing does not need the preparation for data transmission. The indexing method needs the preparation for 2 time slots. For the IP addressing method, data rate is small because the size of header is large. In Table 2, the data transmission rate is the same for Packed-4 (P4) because data transmission size is the same for IP fragmentation based on the MTU (Message Transfer Unit) value of Link-16.

Table 2. Data Rate for IP packet Transmission

	Address input (IP and UDP)	**Indexing scheme (Connection setting)**
Preparation process	None	2 time slots
Header Length	45 bits	18 bits
Data rate	16.26 kbps (ST) 40.64 kbps (P2) 97.54 kbps (P4)	24.38 kbps (ST) 48.77 kbps (P2) 97.54 kbps (P4)

5 Conclusion

Header compression is necessary for IP over a tactical data link such as Link-16. This paper proposed a novel header compression method for IP over Link-16. And the

Monolithic Message was proposed for the efficient IP packet transmission of the tactical data link message. An indexing method was used for transmitting typical IP packet. In addition, the function of transmitting ACK to the compressed header was added in order to complement the disadvantage of the UDP application.

Compared to existing Link-16, 90.7% for IP packet transmission and maximum 95.8% for tactical data link message can be achieved by the proposed one.

As a further study, a research including physical layer issues for efficient IP application over Link-16 will be considered.

Acknowledgments. The research is supported by the Agency for Defense Development (ADD) of Korea.

References

1. US DoD, Network Centric Warfare Report To Congress (July 2001)
2. Jee, S.-B.: An Analysis for Efficient Appliance of TDL Protocol and IP on Link-K System. Agency of Defense Development (in Korean)
3. Wilson, W.J.: Applying Layering Principles to Legacy Systems; Link-16 as a case study. In: IEEE MILCOM 2001, pp. 526–531 (2001)
4. Understanding Link-16: A Guidebook for New User, Northrop Grumman Corporation, San Diego, C.A. (September 2001)
5. EFFNET, An Introduction to IP Header Compression, Effnet White Paper (February 2004)
6. Degermark, M., Nordgren, B., Pink, S.: IP Header Compression. RFC 2507 (Feburary 1999)
7. Casner, S., Van Jacobson.: Compressing RTP/UDP/IP Headers for Low-Speed Serial Links. RFC 2508 (February 1999)
8. Bormann, C.(ed.): Robust Header Compression (ROHC). In: RFC 3095 (June 2001)
9. Woo, H.J., Kim, J.Y., Lee, M.J.: Performance analysis of Robust Header Compression over Mobile WiMAX. In: ICACT 2008, pp. 311–313 (2008)

A Frequency-Based Algorithm for Workflow Outlier Mining

Yu-Cheng Chuang[1], PingYu Hsu[1], MinTzu Wang[2], and Sin-Cheng Chen[1]

[1] Department of Business Administration, National Central University
tocasper@hotmail.com, pyhsu@mgt.ncu.edu.tw,
93441024@cc.ncu.edu.tw
[2] Department of Information Management,
Technology and Science Institute of Northern Taiwan
mtwang@tsint.edu.tw

Abstract. The concept of workflow is critical in the ERP (Enterprise Resources Planning) system. Any workflow that is irrationally and irregularly designed will not only lead to an ineffective operation of enterprise but also limit the implementation of an effective business strategy. The research proposes an algorithm which makes use of the workflow's executed frequency, the concept of distance-based outlier detection, empirical rules and Method of Exhaustion to mine three types of workflow outliers, including less-occurring workflow outliers of each process (abnormal workflow of each process), less-occurring workflow outliers of all processes (abnormal workflow of all processes) and never-occurring workflow outliers (redundant workflow). In addition, this research adopts real data to evaluate workflow mining feasibility. In terms of the management, it will assist managers and consultants in (1) controlling exceptions in the process of enterprise auditing, and (2) simplifying the business process management by the integration of relevant processes.

Keywords: ERP, BPM, Workflow mining, Data mining, Outlier detection.

1 Introduction

The current highly competitive business environment is comprised of global markets, changeable demands of customers and a ferocious competition. Enterprises must strengthen inside operations to maintain a competitive advantage, with a view of making an enormous profit.

The whole business operating strategy ought to be firmly grounded on the utmost maximization of the competence in profiting and increasing the revenue of the enterprise and the satisfaction of the customers. In order to fulfill this, the position of the customers and the ways to gratify them shall be taken into consideration in the process of business organization. Meanwhile, these elements need to be integrated into the whole business operating processes from customers to suppliers [6]. In view of this, many enterprises use information systems like ERP、SCM、CRM、B2B and WfMS to execute or integrate all business processes. These information systems will keep track of the processing details, which are presented as system logs.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 191–207, 2010.

Workflow management system (WfMS), such as Staffware, IBM MQSeries, COSA, based on the definition of WfMC (The Workflow management Coalition), is a system to define, manage and implement workflow. The design of the system must present the logic of workflow completely.

Similar to the concept of a workflow management system, Business Process Management (BPM) is defined as the software and tool to establish and implement the business workflow model by dint of the definition and integration of any necessary personnel, system, application and applied unit, according to the Butler Group [20]. Four major parts are included in BPM: (1) The analysis and construction of the workflow,(2) workflow management,(3) the application and integration of the business, and (4) the surveillance and management of the workflow.

In addition, the implementation of ERP is deemed as one of the pivotal issues. With reference to the example of ASAP (i.e. the implementation method of SAP), the implementation can be divided into five stages, including (1) project preparation, (2) business blueprint, (3) system realization, (4) system final preparation and (5) system co-live & continuous improvement. Consequently, workflow outlier mining is instrumental in terms of the execution of the analysis and construction of the workflow, which in turns helps BPM impeccably integrate the business processes into the information system and improves the efficiency of the execution of ERP and modify the system variables to fine-tune scenarios in the stages of business blueprint and system realization beforehand.

In terms of the management, it can assist managers and consultants in:

1. Managing the exception effectively in the process of internal control and auditing.
2. Simplifying business operating processes by dint of the integration of pertinent processes. For instance, the installation of the BPM system is conducive not only to the analysis of the workflow and the execution of the construction, with a view to producing optimal integration of business processes and the information system, but also to adjust system variables in the process of the installation of ERP by the omission of unnecessary steps of workflow to optimize the system.

By the way, most previous researches pertaining to workflow are focusing on how to reconstruct the workflow process by system log and very few of them are to mine workflow outlier. According to the motivation for this research, one objective of this research is to excogitate an algorithm suitable for every system log for mining less-used or never-used business processes. The features of system logs will be the primary focus to probe into the workflow outlier by means of executed process frequency and statistical methods. In addition, the Method of Exhaustion will be adopted to find out the never-occurring workflow outliers.

The remainder of this paper is organized as follows. Section 2 is mainly with literature review, offering references pertinent to workflow mining and outlier detection. The process model and the algorithm for mining adopted in this research will be elucidated in Section 3, which demonstrates the methodology and the algorithm of this research. The implementation of the algorithm as a rudimental system will be conducted in Section 4 to assess the accuracy of this algorithm, while the results of the research and the direction of future researches will be summarized in Section 5.

2 Literature Review

The pertinent references, including (1) workflow mining and (2)outlier detection. The research of Process Discovery has always been the most prevalently discussed in the realm of workflow mining (see Fig.1).

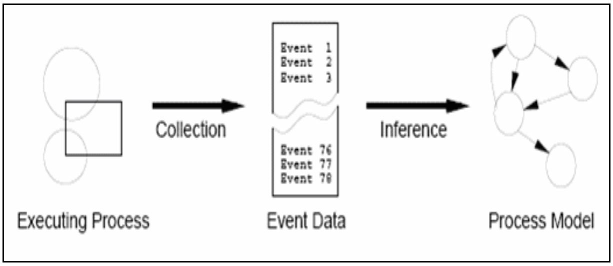


Fig. 1. Process Discovery

2.1 Workflow Mining

In Cook & Wolf's research [13], they suggest three methods for process discovery, inclusive of neural networks, as the purely algorithmic approach and the Markov approach the authors consider being the most promising. On one hand, the purely algorithmic approach establishes a finite state machine (FSM) to connect any similar recurrence in the future, which produces the method for the final process model. On the other hand, the Markov approach adopts a mixture of algorithmic and statistical methods, having the ability to deal with noise. This research is conducted within a sequential behavior.

Cook and Wolf extend their study to concurrent processes [14]. They devise specific metrics (entropy, event type counts, periodicity, and causality) and use these metrics to discover models out of event streams. Yet, they do not provide any approach productive of specific process models.

The idea of the application of process mining to the real in the context of workflow management was first introduced in [19], which is grounded on workflow graphs under the inspiration of workflow products like IBM MQSeries workflow (formerly known as Flowmark) and InConcert. In this paper, two problems are defined. The first problem is to find the occurrence of events that a workflow graph generates in a given workflow log. The second problem is to find the definitions of edge conditions. A concrete algorithm is served as a means to tackling the first problem. The approach differs considerably from the others by reason of the nature of workflow graphs; therefore, it is needless to identify the nature (AND or OR) of merges and splits. Although in terms of the handling of a partial situation of recursion, Method of Exhaustion is adopted and the potential graphs are adjusted in this paper, the process model is incompletely presented.

Schimm [5] has developed a mining tool suitable for discovering hierarchically structured workflow processes. This requires the maintenance of the equilibrium of all splits and merges.

Herbst and Karagiannis also address the issue of process mining in the context of workflow management by means of an inductive approach. The approach presented respectively in [8][9][10][11][12] allows for concurrency, using stochastic task graphs as an intermediate representation and generating a workflow model described in the ADONIS modeling language. In the process of induction, task nodes are merged and divided in order to discover the underlying process. A notable difference from other approaches is that the same task node can appear multiple times in the workflow model. In [1] & [2], a heuristic approach rather than simple metrics is used to construct so-called "dependency/frequency tables" and "dependency/frequency graphs".

Among the researches pertaining to workflow mining, most of them concentrate on the process discovery of workflow mining. However, in [18], the researchers mine a process model able to adjust configures by each activity's execution frequency. It is pointed out in this research that the application of the concept of workflow mining to Information System such as ERP, CRM and SCM will contribute abundant commercial benefit to the enterprise and fortify its competitiveness. Hence, how to detect and mine the workflow outlier is another important focus for workflow mining.

2.2 Outlier Detection

At present, the bulk of studies relevant to outlier detection are categorized as a part of statistics. At this time an "outlier", or "noise", receives no categorically unanimous definition, Hawkins proffers a definition inclusive thoroughly of the characteristics and the pith of an outlier, indicating that, "An outlier is an observation that it was generated by a different mechanism [7]." Outlier detection is one of the fundamental issues in data mining, especially in fraud detection, network intrusion detection, network monitoring, etc. The traditional methods of outlier detection are classified as follows:

(1) Distribution-based methods in this category were previously conducted by the statistics community. Some standard distribution models (e.g. normal) are employed in these methods and those points that deviate from the model as outliers are flagged. Recently, Yamanishi, Takeuchi and Williams [23] used a Gaussian mixture model (GMM) to present that normal behaviors and each datum is scored based on changes in the model. A high score likely manifests a high possibility as an outlier.

This approach has been in combination with a supervision-based learning approach to obtain general patterns for outliers [22]. In view of arbitrary data sets without any prior knowledge of the distribution of points, conducting expensive tests to determine which model fits the data best, if any, is ineluctable.

In 2002, Yamanishi and Takeuchi suggest an integrated framework to cope with both of them on the basis of the theory of the on-line learning of non-stationary time series [21].

(2) Deviation-based techniques ascertain outliers by inspecting the characteristics of objects and deem an object that diverges from these characteristics as an outlier [3].

(3) Distance-based methods are originally proposed by Knorr and Ng [15][16][17]. This notion generalizes many ideas from the distribution-based approach and renders better picture of computational complexity.

(4) The Density-based method is proposed by Breunig, Kriegel, Ng, and Sander [4]. It relies on the local outlier factors (LOF) of each point in accordance with the local density of its neighborhood. From the typical use, points with a high LOF are flagged as outliers.

In conclusion, as far as this research is concerned, the activity's execution frequency and a distance-based method for Workflow Outlier Mining will be adopted. After discovering the workflow outlier, this research also makes use of process reconstructing redundant process from never occurring workflows.

3 Methodology

The concepts pertinent to workflow will be introduced in this section. In addition, the algorithm for workflow outlier mining will be designed in accordance with the characteristics of workflow.

3.1 Related Workflow Definitions

According to the definition of Workflow from Workflow Management Coalition (WfMC), here are relevant concepts of workflow:

Definition 1: Workflow is also known as Process, which is to connect predefined rules and transmit the document, information or task among participants of the process.

Definition 2: Activity is also known as "Job" or "Task", which is used to describe one task in a workflow. In logical terms, it can be seen as the smallest complete task. Not all of the activities, which can be divided into either manual or automatic activity, are executed automatically. The manual activity is indicative that the workflow execution should be executed under the activity participant's assistance or supervision. The automatic activity means that the activity is executed by the trigger program and works independently.

Definition 3: Instance is the actual situation of workflow execution. It is feasible that many Instances are running simultaneously without interfering with each other.

Generally speaking, workflow can be judged from two points of view: Design Time and Run Time. From the perspective of Design Time, all the workflows are predefined. Take Fig.2 for example. Many Processes (P_1, P_2, P_3…P_n) are involved in a system, and each process is comprised of many Activities (A_1, A_2, A $_3$…A_n), each of which has its own type. Take the Purchase Order Process for example. A_1 may be the START type, indicative of the activation of a purchase order; A_2 may be the WEB type and it means a purchaser list that contains all required materials; A_3 may be the XOR type, representing IS checks the rationality of material specification; A_4 maybe the SUBFLOW type, standing for the following procedure of purchase order; A_5 may be the END type, representing the end of this process.

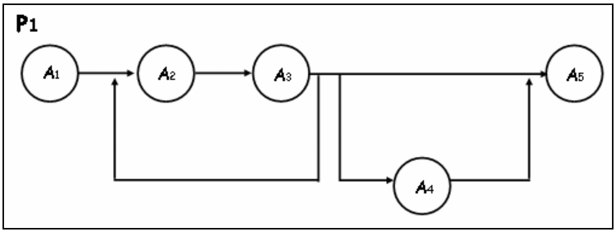


Fig. 2. The Concept of Process

From the viewpoint of Run Time, the record of execution named "Instance", which is comprised of WorkItem. As shown in Fig.3, Process P_1 has been executed for four times in a real situation, leaving the execution record Instance I_1, I_2, I_3, $I_{4.}$ The actual execution situation is demonstrated as follows: I_1's execution sequence is W_1, W_2, W_3, W_5; I2's execution sequence is W_1, W_2, W_3, W_4, W_5; I_3's execution sequence is W_1, W_2, W_3, W_2, W_3, W_5; I_4's execution sequence is W_1, W_2, W_3, W_2, W_3, W_4, W_5.

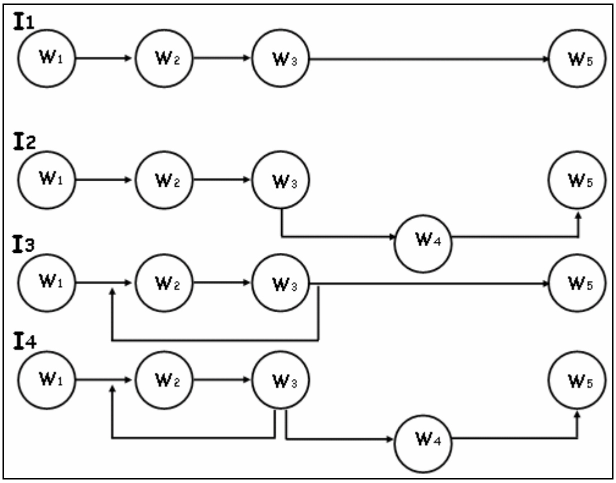


Fig. 3. The Concept of Instance

3.2 The Algorithm for Workflow Outlier Mining

In this section, the application of the concepts of workflow outlier to the workflow will be conducted, and the algorithm for workflow outlier mining will be contrived in accordance with the characteristics of workflow. The task for workflow outlier mining will be divided into two parts: One is to search for the less-frequent abnormal workflow appearing in the system, and the other aims at the never-executed redundant workflow in the system.

According to the Empirical Rule of Statistics, the probability of data located at the range [X-S, X+S] is 68%; the probability of data located at the range [X-2S, X+2S] is 95%; the probability of data located at the range [X-3S, X+3S] is 99.7%. Based on the motivation of the research mentioned previously, it will provide consultants and managers with valuable information if less-occurring workflow (also called Abnormal Workflow) can be mined from Instance, which represents the actual execution situation.

Consequently, the concept of FPOF (Frequent Pattern of Factor) [24] will be ameliorated into FIOF (Frequent Instance of Factor), which is suitable for workflow outlier mining. FIOF is used to measure the outlier's degree of Instance and able to form the data range in combination with Empirical Rule, and finally workflow outlier will be mined.

Abnormal Workflow can be divided into two types. The first type is less-occurring instance in each process. Instance's Process Activities sequence is Abnormal Workflow. Take Fig.4 for example. In the actual execution records, Instance I1's frequency is twice. Apparently it is a kind of less-occurring workflow, and its Process Activities sequence (A1, A2, A3, A8) is the workflow outlier.

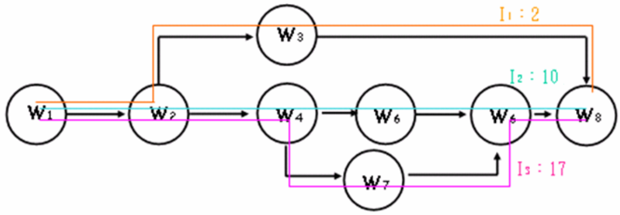


Fig. 4. Less-occurring workflow

In terms of the first type of workflow outlier, an algorithm called "Abnormal WMe" (Abnormal Workflow Mining for each process) is designed, and the procedure is shown as follows:

(1) Take all average numbers of the Instance (in each Process) support as minimum support (min_support).

(2) Find out the number of the Instance support larger than the min_support, which is called frequent instance in each Process. The number of frequent instance is called ||FIS (P, min_support)||. FIS (Frequent Instance Set) means the set of instance in which frequency is higher than the threshold of min_support.

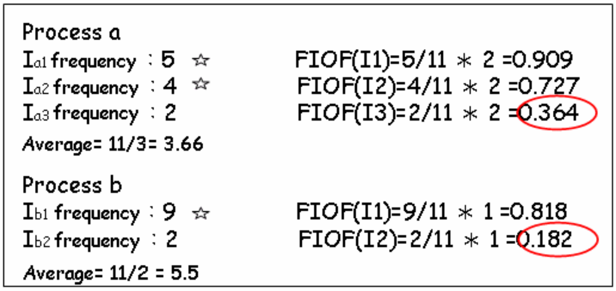
(3) Calculate each Instance's FIOF (Frequent Instance of Factor).

FIOF(i) = support(i)*||FIS(P, min_support)||

(4) Use the Empirical Rule to calculate the probability range of FIOF. Those Instances' FIOF numbers located at the left side of range are Abnormal Workflow.

The point of this algorithm, is that after the calculation of the frequency of each Instance, the instance support's average number will be viewed as min_support, which represents the trend this data set focuses on. If an Instance's support is lower than min_support, it is likely to be an outlier. Then calculate each Instance's FIOF, and use the Empirical Rule to find out relatively less-occurring Instance, of which Process Activities sequence is Abnormal Workflow.

In step 3- the calculation of FIOF, the purpose of being multiplied by ||FIS(P, min_support)|| is to emphasize Instance's degree of outlier if its Process has fewer frequent instances. As shown in Fig.5, Process a and Process b both have an Instance's support that is 2/11=0.182. It is probably inferred that both of them are outliers. But it is observable that in Process b, Ib2's frequency (2) is far less than Ib1's frequency (9). In Process a, though Ia3's frequency is also 0.182, it is not far different from other Instances Ia1 and Ia2. Hence, being multiplied by ||FIS(P, min_support)|| can underscore the degree of outlier, and as a result, Ib2 has higher probability to be outlier than Ia3 (0.182<0.364).



Process a

I_{a1} frequency : 5 ☆ FIOF(I1)=5/11 ∗ 2 =0.909

I_{a2} frequency : 4 ☆ FIOF(I2)=4/11 ∗ 2 =0.727

I_{a3} frequency : 2 FIOF(I3)=2/11 ∗ 2 =0.364

Average= 11/3= 3.66

Process b

I_{b1} frequency : 9 ☆ FIOF(I1)=9/11 ∗ 1 =0.818

I_{b2} frequency : 2 FIOF(I2)=2/11 ∗ 1 =0.182

Average= 11/2 = 5.5

Fig. 5. The Example of How ||FIS(P, min_support)|| Affects FIOF

The detailed algorithm AbnormalWMe is displayed as Fig 6:

The second type of Workflow Outlier is the less-occurring Process in all Processes relatively. The procedure is shown as follows:

(1) Sort all Processes by their total Instance number, and take all Process' support average as min_support.

Use the Empirical Rule to find out the Workflow Outlier located beyond the range.

Algorithm 1. AbnormalWMe (abnormal workflow mining for each process)

```
Input： P, all Process' set,  P= (p1,p2,p3…pi…pn) ;
        I(pi),each Process' Instance's set,  I(pi)=(i1,i2,i3…ii…in) ;
          S, all Instance's support's set,  S=(s1,s2,s3…si…sn) ;
Output： Abnormal workflow set AW,

for each Process pi do
 min_support(pi)= Average number of Support;
 //Find each Process' min_support。
 FIS_number(pi)
 //Find how many frequent instances are larger than min_support，||FIS(P,min_support)||。
```

Fig. 6. The AbnormalWMe Algorithm

```
{
 int number=0;
  for (i=1; i<=n; i++)
   if (si >= min_support) FIS_number=FIS_number+1;
   else return FIS_number;
  return FIS_number;
 //Search from the first instance to the last one. If the instance support is larger than
   min_support, add the number.
}

AbnormalWMe(P)
//Mine less-occurring workflow in each Process, and its Process Activities
sequence is the workflow outlier
{
  for each Instance ii do
     FIOF(ii)= si * FIS_number(pi)
      //Calculate each instance's FIOF($i_i$)=support(i) * ||FIS(P,min_support)||
     Calculate X(arithmetic mean of FIOF(I))and S(standard deviation of FIOF(I))
       //Calculate all Instances' FIOF average X and standard deviation S
   let lower_bound=X-S;
   //According to the Empirical Rule, there is 68% data which lower bound is X-S
   for each FIOF(ii) do selection_sorting();
   //Use selection sorting to sort all Instances from small FIOF to large FIOF
      for (ii=1; ii< n; ii ++)
       if (FIOF(ii)<=lower_bound) return AW[ii] ;
       else return 0;
      return 0;
   //Start from the smallest FIOF. If FIOF is lower than lower_bound, return this
      Instance's Process Activities sequence. These left range outliers are
      Abnormal Workflow.
}
```

Fig. 6. (*continued*)

3.3 Redundant Workflow Mining

The other type of Workflow Outlier consists of those Processes' Instances that never-occur in system logs. It is possibly because there are many probably-occurring sequences (Instance) in a Process, while some sequences are not suitable in a real scenario. Hence, some sequences are never-occurring. Regardless of BPM Project or ERP Implementation, finding out these never-occurring sequences will improve the enterprise's operating efficiency.

The concept is to use Method of Exhaustion to find out every Process' potential execution sequence and subtract with Instance (Process Activities sequence that really occur) to mine the never-occurring Workflow.

As demonstrated in Fig. 7(a) through Fig. 7(c):

(1) The first graph is a Process’ original model. The nearest split type Activity shall be searched for from the last Activity.

(2) In the second graph, the nearest split type Activity E is found. Start from E to build Activity Path. Use E[2] to present two paths of E: (EJLN) and (EKMN).

(3) In the third graph, the next nearest split type Activity G is discovered. Start from G to build Activity Path. Use G[3] to present three paths of G: (GDE[2])and (GHIKMN). It means (GDEJLN), (GDEKMN) and (GHIKMN).

(4) In the fourth graph, the last split type Activity A is detected, which is the start of the Process. Start from A to build Activity Path, and thereby use A[5] to present five paths of A: (ABCDE[2]) and (AFG[3]); that is, (ABCDEJLN), (ABCDEKMN), (AFGDEJLN), (AFGDEKMN), and (AFGHIKMN). Because Activity A is this Process’ start type Activity, the five paths are all possible Activity sequence. This means Set(P)={ (ABCDEJLN) 、(ABCDEKMN) 、(AFGDEJLN)、(AFGDEKMN)、(AFGHIKMN)}

(5) In the fifth graph is the Instance (ever-occurred Process Activity sequence) in system logs.

    Instance(P)={ (ABCDEJLN) 、(AFGDEKMN) }

(6) The sixth graph is representative of the subtract of Set(P) and Instance(P), which is the never-occurring Workflow in the system. It includes { (ABCDEKMN) 、(AFGDEJLN) 、(AFGHIKMN) }

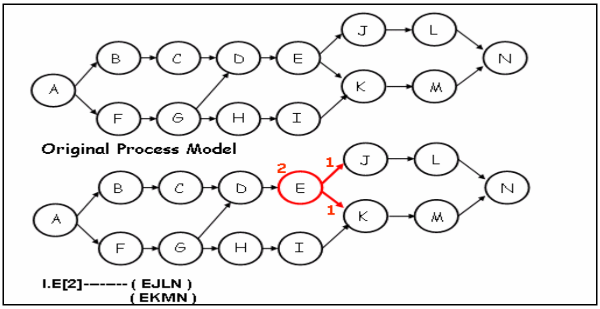
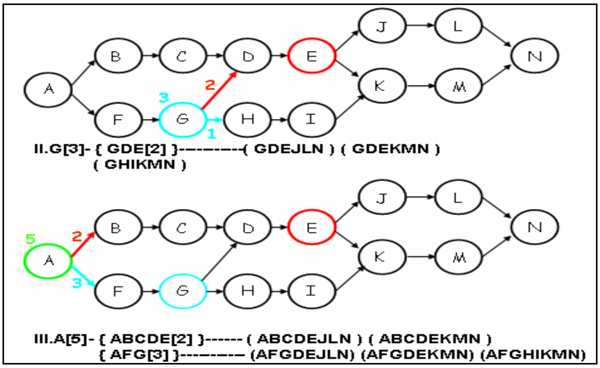


Fig. 7a. The Procedure of Mining Redundant Workflow



A
B
C
D
E
F
G
H
I
J
K
L
M
N
3
2
1
II.G[3]- { GDE[2] }---------(GDEJLN) (GDEKMN)
(GHIKMN)
5
2
3
III.A[5]- { ABCDE[2] }------(ABCDEJLN) (ABCDEKMN)
{ AFG[3] }--------- (AFGDEJLN) (AFGDEKMN) (AFGHIKMN)

Fig. 7b. The Procedure of Mining Redundant Workflow

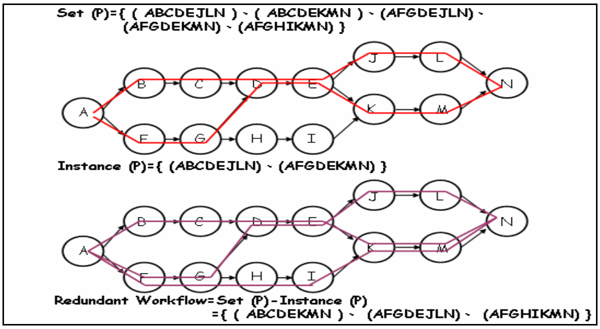

Set (P)={ (ABCDEJLN) ､ (ABCDEKMN) ､ (AFGDEJLN) ､ (AFGDEKMN) ､ (AFGHIKMN) }
A
B
C
D
E
F
G
H
I
J
K
L
M
N
Instance (P)={ (ABCDEJLN) ､ (AFGDEKMN) }
A
B
C
D
E
F
G
H
I
J
K
L
M
N
Redundant Workflow=Set (P)-Instance (P)
={ (ABCDEKMN) ､ (AFGDEJLN) ､ (AFGHIKMN) }

Fig. 7c. The Procedure of Mining Redundant Workflow

4 Experiments and the Evaluation

This section is divided into two parts. The methods for the design of experiments will be elucidated in the first part, while the results of experiments will be further explicated and analyzed in the second part.

4.1 The Design of Experiments

This experiment is designed to utilize the real Workflow System Log to validate the feasibility of algorithm, with a view to mining three discrepant types of Workflow Outliers, including less-occurred Workflow in each Process, less-occurred Workflow in all Processes, and never-occurred Workflow in each Process.

The procedure of the experiment is shown as follows:

(1) Dump workflow system logs into the database.
(2) Calculate each Process' helpful figures like the support of each Process and that of each Instance.
(3) Find out all Workflow Outliers in advance to compare the accuracy of algorithm according to the respective definitions of three Workflow Outliers.
(4) Put Workflow Log into the experimental system to mine Workflow Outliers.
(5) Compare the result with identified Workflow Outliers to evaluate the accuracy of algorithm.

In respect to the real data, it emanates from the backup archive stored in certain domestic manufacturing company's workflow management system Log. The record Instance contains the information over six months (from July, 2006 to December, 2006). There are twenty-five Predefined Process Models, only nine ever-occurred Processes exist. It is probably because the data source records for six months only while many Processes still continue running. There are about thousands of instances, however; only 68 among them are finished. It means the other Workflows are not finished but still running. Nevertheless, in this research, only Complete Workflows are focused on to mine Workflow Outlier. As a consequence, the data amount is tenfold to six hundred and eighty Complete Workflows which generated by discrete distribution with proportional probability to simulate actual data volume. Based on the confidential contract with the cooperative company, all the Workflow names have been changed.

Besides, to make the algorithm more efficient, adjusted metadata is shown as follows:

✓ **Process(DEFINITIONID、ACTIVITYNAME、ACTIVITYTYPE、SPLIT_NEXT)**

DEFINITIONID represents each Process that includes many Activities; ACTIVITYNAME stands for each Activity's name▯ ACTIVITYTYPE means each Activity Type, including START, SPLIT and END▯ SPLIT_NEXT is a newly added field and data, which is used to record this SPLIT type Activity's next connected ACTIVITYNAME. This data structure will conduce to the ideal performance of the algorithm.

✓ **Instance(INSTANCEID、ACTIVITYNAME、ACTIVITYTYPE)**

INSTANCEID is indicative of each Instance; ACTIVITYNAME and ACTIVITYTYPE record Activity information, where the Instance is executed through.

From the observation of these Instances, some Workflow executed loops appear more than once. The objective of Workflow Outlier Mining is to find out Workflows that don't happen frequently. If any execution frequency of loop appears more than once, it will be classified as the same one. As shown in Fig.8, I_3 and I_4 of this Process are executed twice and also three times separately in (W_2, W_3). These situations are classified as I_2, the loop that is executed only once.

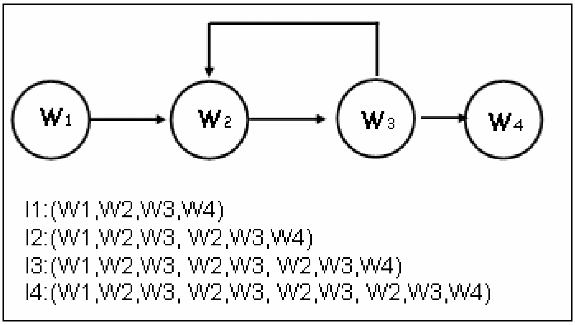

W1
W2
W3
W4
I1:(W1,W2,W3,W4)
I2:(W1,W2,W3, W2,W3,W4)
I3:(W1,W2,W3, W2,W3, W2,W3,W4)
I4:(W1,W2,W3, W2,W3, W2,W3, W2,W3,W4)

Fig. 8. How to deal with Loop Iteration

Furthermore, in this research, the Process model representative of the connection of Activities has been comprehended beforehand, which will assist the algorithm in building Process Activities sequence in the process of dealing with Redundant Workflow.

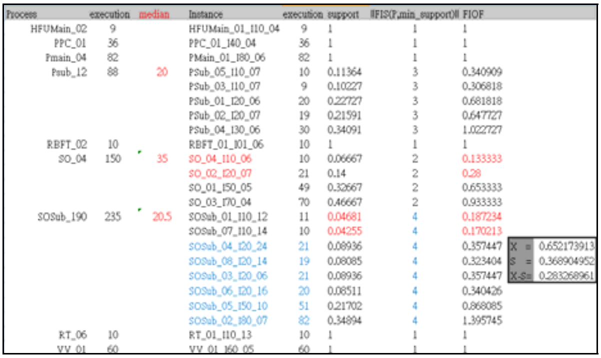
4.2 The Result and Analysis of Experiments

Experiment 1- AbnormalWMe

Experimental Object: Mine the less occurred Instance in each Process, of which Process Activities sequence is deemed as Abnormal Workflow.

Calculate each Instance's support, and use the support average number as min_support. Calculate ||FIS(P, min_support)|| (how many Instance's supports are larger than min_support), and calculate FIOF. All FIOF's average number X is 0.652; standard deviation S is 0.369; X-S is 0.283. Consequently, Workflow Outliers should be those FIOF smaller than 0.283. The extra Outliers are the Process Activities sequence of these two Instances- SO_02 and SO_04 as Fig.9 shows.

If without FIFO, we may get different results such as SOSub_01, SOSub_07. After the observation of these Instances, Process SO's Instance execution frequencies are 10, 21, 49 and 70. It is perceived clearly that 10 and 21 Instance execution frequencies are significantly lower than 49 and 70.



Process	execution	median	Instance	execution	support	‖FIS(P,min_support)‖	FIOF		
HFUMain_02	9		HFUMain_01_110_04	9	1	1	1		
PPC_01	36		PPC_01_140_04	36	1	1	1		
Pmain_04	82		PMain_01_180_06	82	1	1	1		
Psub_12	88	20	PSub_05_110_07	10	0.11364	3	0.340909		
			PSub_03_110_07	9	0.10227	3	0.306818		
			PSub_01_120_06	20	0.22727	3	0.681818		
			PSub_02_120_07	19	0.21591	3	0.647727		
			PSub_04_130_06	30	0.34091	3	1.022727		
RBFT_02	10		RBFT_01_101_06	10	1	1	1		
SO_04	150	35	SO_04_110_06	10	0.06667	2	0.133333		
			SO_02_120_07	21	0.14	2	0.28		
			SO_01_150_05	49	0.32667	2	0.653333		
			SO_03_170_04	70	0.46667	2	0.933333		
SOSub_190	235	20.5	SOSub_01_110_12	11	0.04681	4	0.187234		
			SOSub_07_110_14	10	0.04255	4	0.170213		
			SOSub_04_120_24	21	0.08936	4	0.357447	X =	0.652173913
			SOSub_08_120_14	19	0.08085	4	0.323404	S =	0.368904952
			SOSub_03_120_06	21	0.08936	4	0.357447	X-S=	0.283268961
			SOSub_06_120_16	20	0.08511	4	0.340426		
			SOSub_05_150_10	51	0.21702	4	0.868085		
			SOSub_02_180_07	82	0.34894	4	1.395745		
RT_06	10		RT_01_110_13	10	1	1	1		
VV_01	60		VV_01_160_05	60	1	1	1		

Fig. 9. The result of AbnormalWMe

Experiment 2- AbnormalWMa

Experimental Object: Mine less-occurred Process in all Process set.

Finding out the less-occurred Process in all Process set is essential, whereas in the whole twenty-five Process Models, only nine ever-occurred Processes exist. This situation will give rise to the bulk of the support numbers that are 0 in the data set. In terms of the frequency-based concept, it will be unproductive of any result.

To supplement the insufficiency of data volume, the random number table is used to simulate all twenty-five Process' Instance numbers. The twenty-five Process' execution frequencies are shown as follows: (CRFT:292,AT:803,EXF:518, HFS:902,MT:925,RP:319,RPI:037,RPIA:407,RCFT:109,SDAI:266,SCTS:275,SPIPCT:784,SOC:340,TT:512,TTT:012,UNT:382,RV:082,HFUMain:554,RBFT:772,RT:205, VV:168,Pmain:129,Psub:031,SO:343,SOsub:625).

After the calculation of the average number X and deviation standard S, the lower bound of support 0.01 is obtained. If the Process' support is lower than 0.01, it is considered as the less-occurred Process in all Processes. The result is displayed as follows: TTT, Psub, RPI, RV, RCFT and Pmain. Although data volume is insufficient, by means of the use of Random Numbers to simulate the execution frequency, it is feasible to validate the algorithm's accuracy.

Experiment 3- RedundantWM

Experimental Object: Mine the Process Activity sequence which is never executed. Find every Process' possible execution sequence in advance. Then, follow the procedure of section 3.3 to find Redundant Workflow. Besides, only nine Process' possible execution sequences are required to be found in this experiment because other twenty Processes are never-occurred.

Before reading data into the system is conducted, a special adjustment for Activity TR14 of Process RT_06 (Fig.10) needs to be made. Activity TR14's ACTIVITYTYPE is converted into SPLIT. SPLIT_NEXT is RT13 and RT8. The reason for this adjustment is that the path from RT7 to RT 13 not only is (RT7,8,9,10,11,12,13), (RT7,8,9,12,13) and (RT7,14,13), but also has the parallel path (RT7, 14,8,9,10,11,12,13) and (RT7,14,8,9,12,13).

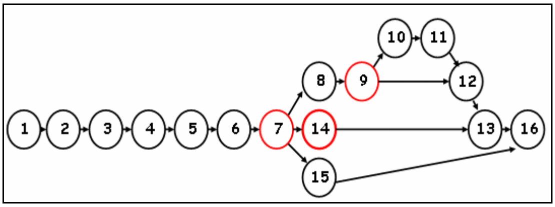


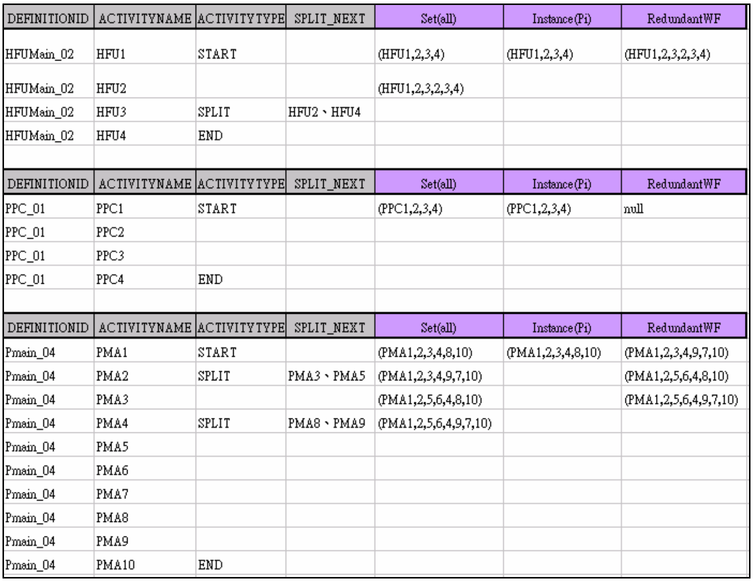
Fig. 10. RT_06 Process Model

The result of Process HFUMain_02, PPC_01, Pmain_04 is demonstrated in Fig.11, while the result of RT_06 and VV_01, Process Psub_12, RBFT_02 and SO_04 is similar.

The left side of figures is every Process' information, while the right side is the result of running algorithm. Set (all) represents sequences of all possible Process Activities. Instance(Pi) is ever-executed Workflow. RedundantWF stands for all the Workflows which are never-occurred, and null indicates zero existence of never-executed Workflow.

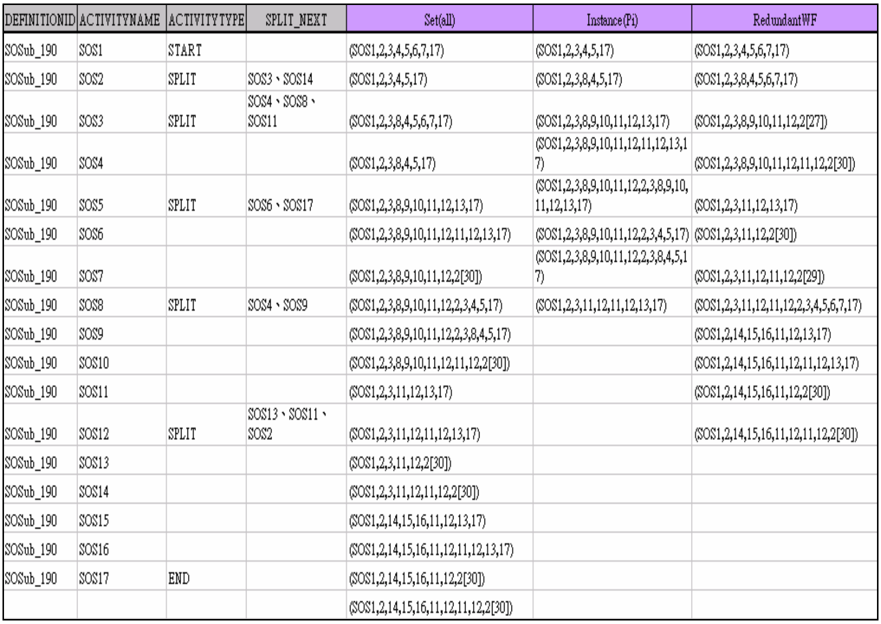
Fig.12 shows the result of Process SOSub_190, which has 190 possible paths. Some Activity SOS2 is presented as SOS2(30), which means thirty repeated execution paths appear repeatedly. According to Process SOSub_190's Process Model, there are sixteen possible paths. However, the algorithm only takes one loop iteration into consideration. If the execution path needs to be repeated more then once, there are only 30(15+15) ways to finish the Process. RedundantWF represents all the Workflows which never-occurred.

Experimental Analysis: In each Process' DEFINITIONID, the number indicates how many possible paths in each Process are calculated in advance; for example, SOSub_190 means Process SOSub has 190 possible execution paths. Set(all) stands for all the possible execution paths calculated by the algorithm.



DEFINITIONID	ACTIVITYNAME	ACTIVITYTYPE	SPLIT_NEXT	Set(all)	Instance(Pi)	RedundantWF
HFUMain_02	HFU1	START		(HFU1,2,3,4)	(HFU1,2,3,4)	(HFU1,2,3,2,3,4)
HFUMain_02	HFU2			(HFU1,2,3,2,3,4)		
HFUMain_02	HFU3	SPLIT	HFU2、HFU4			
HFUMain_02	HFU4	END				
DEFINITIONID	ACTIVITYNAME	ACTIVITYTYPE	SPLIT_NEXT	Set(all)	Instance(Pi)	RedundantWF
PPC_01	PPC1	START		(PPC1,2,3,4)	(PPC1,2,3,4)	null
PPC_01	PPC2					
PPC_01	PPC3					
PPC_01	PPC4	END				
DEFINITIONID	ACTIVITYNAME	ACTIVITYTYPE	SPLIT_NEXT	Set(all)	Instance(Pi)	RedundantWF
Pmain_04	PMA1	START		(PMA1,2,3,4,8,10)	(PMA1,2,3,4,8,10)	(PMA1,2,3,4,9,7,10)
Pmain_04	PMA2	SPLIT	PMA3、PMA5	(PMA1,2,3,4,9,7,10)		(PMA1,2,5,6,4,8,10)
Pmain_04	PMA3			(PMA1,2,5,6,4,8,10)		(PMA1,2,5,6,4,9,7,10)
Pmain_04	PMA4	SPLIT	PMA8、PMA9	(PMA1,2,5,6,4,9,7,10)		
Pmain_04	PMA5					
Pmain_04	PMA6					
Pmain_04	PMA7					
Pmain_04	PMA8					
Pmain_04	PMA9					
Pmain_04	PMA10	END				

Fig. 11. The Result of RedundantWM (1)



DEFINITIONID	ACTIVITYNAME	ACTIVITYTYPE	SPLIT_NEXT	Set(all)	Instance(Pi)	RedundantWF
SOSub_190	SOS1	START		(SOS1,2,3,4,5,6,7,17)	(SOS1,2,3,4,5,17)	(SOS1,2,3,4,5,6,7,17)
SOSub_190	SOS2	SPLIT	SOS3、SOS14	(SOS1,2,3,4,5,17)	(SOS1,2,3,8,4,5,17)	(SOS1,2,3,8,4,5,6,7,17)
SOSub_190	SOS3	SPLIT	SOS4、SOS8、SOS11	(SOS1,2,3,8,4,5,6,7,17)	(SOS1,2,3,8,9,10,11,12,13,17)	(SOS1,2,3,8,9,10,11,12,2[27])
SOSub_190	SOS4			(SOS1,2,3,8,4,5,17)	(SOS1,2,3,8,9,10,11,12,11,12,13,17)	(SOS1,2,3,8,9,10,11,12,11,12,2[30])
SOSub_190	SOS5	SPLIT	SOS6、SOS17	(SOS1,2,3,8,9,10,11,12,13,17)	(SOS1,2,3,8,9,10,11,12,2,3,8,9,10,11,12,13,17)	(SOS1,2,3,11,12,13,17)
SOSub_190	SOS6			(SOS1,2,3,8,9,10,11,12,11,12,13,17)	(SOS1,2,3,8,9,10,11,12,2,3,4,5,17)	(SOS1,2,3,11,12,2[30])
SOSub_190	SOS7			(SOS1,2,3,8,9,10,11,12,2[30])	(SOS1,2,3,8,9,10,11,12,2,3,8,4,5,17)	(SOS1,2,3,11,12,11,12,2[29])
SOSub_190	SOS8	SPLIT	SOS4、SOS9	(SOS1,2,3,8,9,10,11,12,2,3,4,5,17)	(SOS1,2,3,11,12,11,12,13,17)	(SOS1,2,3,11,12,11,12,2,3,4,5,6,7,17)
SOSub_190	SOS9			(SOS1,2,3,8,9,10,11,12,2,3,8,4,5,17)		(SOS1,2,14,15,16,11,12,13,17)
SOSub_190	SOS10			(SOS1,2,3,8,9,10,11,12,11,12,2[30])		(SOS1,2,14,15,16,11,12,11,12,13,17)
SOSub_190	SOS11			(SOS1,2,3,11,12,13,17)		(SOS1,2,14,15,16,11,12,2[30])
SOSub_190	SOS12	SPLIT	SOS13、SOS11、SOS2	(SOS1,2,3,11,12,11,12,13,17)		(SOS1,2,14,15,16,11,12,11,12,2[30])
SOSub_190	SOS13			(SOS1,2,3,11,12,2[30])		
SOSub_190	SOS14			(SOS1,2,3,11,12,11,12,2[30])		
SOSub_190	SOS15			(SOS1,2,14,15,16,11,12,13,17)		
SOSub_190	SOS16			(SOS1,2,14,15,16,11,12,11,12,13,17)		
SOSub_190	SOS17	END		(SOS1,2,14,15,16,11,12,2[30])		
				(SOS1,2,14,15,16,11,12,11,12,2[30])		

Fig. 12. The Result of RedundantWM (3)

5 Conclusions and Suggestions

The research proposes an algorithm which makes use of the workflow's executed frequency, the concept of distance-based outlier detection, empirical rules and Method of Exhaustion to mine three types of workflow outliers, including less-occurring workflow outliers of each process, less-occurring workflow outliers of all processes and never-occurring workflow outliers. Besides, this research also adopts real data to evaluate workflow mining feasibility. To sum up, these algorithms can help consultants and managers find workflow outliers and adjust them in anticipation of efficient execution of business processes.

Among three experiments of this research, we suffer from insufficient data volume. Here, suggestions from experiments are proffered below :(1)in order to gain any result worthy of being experimented, a data set that lasts for a long period is required, and (2) attempt to make the algorithm as impeccable and tenable as possible when facing every kind of Process.

In the midst of plentiful methods available to mine outliers, frequency-based and distance-based concepts are applied in this research by virtue of the features of the workflow. In the future, it is likely that other methods will be employed to develop the algorithm for mining workflow outlier. With regards to many the different methods for mining that have been developed, available for us to conclude which method can be conducted in each situation.

In addition, loop iteration is also a critical issue to deal with. In this research, the numbers of iterations of loop are reduced to one, but human judgment is involved. In future research, it is essential to integrate this issue into the algorithm, with a view to reducing any potential inaccuracy.

References

[1] Weijters, A.J.M.M., van der Aalst, W.M.P.: Process mining: discovering workflow models from event-based data. In: Kr€ose, B., de Rijke, M., Schreiber, G., van Someren, M. (eds.) Proceedings of the 13th Belgium–Netherlands Conference on Artificial Intelligence (BNAIC 2001), pp. 283–290 (2001)

[2] Weijters, A.J.M.M., van der Aalst, W.M.P.: Rediscovering workflow models from event-based data. In: Hoste, V., de Pauw, G. (eds.) Proceedings of the 11th Dutch-Belgian Conference on Machine Learning (Benelearn 2001), pp. 93–100 (2001)

[3] Arning, A., Agrawal, R., Raghavan, P.: A linear method for deviation detection in large databases. In: Proceedings of the KDD 1996, pp. 164–169 (1996)

[4] Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: identifying density-based local outliers. In: Proceedings of the SIGMOD 2000, pp. 93–104 (2000)

[5] Schimm, G.: Process Mining, `http://www.processmining.de/`

[6] Gartner Group, `http://www.comwave.com.tw/crm-solution/defi.htm`

[7] Hawkins, D.: Identification of outliers. Chapman & Hall, Reading (1980)

[8] Herbst, J.: A machine learning approach to workflow management. In: Lopez de Mantaras, R., Plaza, E. (eds.) ECML 2000. LNCS (LNAI), vol. 1810, pp. 183–194. Springer, Heidelberg (2000)

[9] Herbst, J., Karagiannis, D.: An inductive approach to the acquisition and adaptation of workflow models. In: Ibrahim, M., Drabble, B. (eds.) Proceedings of the IJCAI 1999 Workshop on Intelligent Workflow and Process Management: The New Frontier for AI in Business, Stockholm, Sweden, pp. 52–57 (August 1999)

[10] Herbst, J., Karagiannis, D.: Integrating machine learning and workflow management to support acquisition and adaptation of workflow models. In: Proceedings of the Ninth International Workshop on Database and Expert Systems Applications, pp. 745–752. IEEE, Los Alamitos (1998)

[11] Herbst, J.: Dealing with concurrency in workflow induction. In: Baake, U., Zobel, R., Al-Akaidi, M. (eds.) European Concurrent Engineering Conference, SCS Europe (2000)

[12] Herbst, J.: Ein induktiver Ansatz zur Akquisition und Adaption von Workflow-Modellen, Ph.D. thesis, Universit€at Ulm (November 2001)

[13] Cook, J.E., Wolf, A.L.: Event-based detection of concurrency. In: Proceedings of the Sixth International Symposium on the Foundations of Software Engineering (FSE-6), pp. 35–45 (1998)

[14] Cook, J.E., Wolf, A.L.: Software process validation: Quantitatively measuring the correspondence of a process to a model. ACM Transactions on Software Engineering and Methodology 8(2), 147–176 (1999)

[15] Knorr, E., Ng, R.: A unified notion of outliers: Properties and computation. In: Proceedings of the KDD 1997, pp. 219–222 (1997)

[16] Knorr, E., Ng, R.: Algorithms for mining distance-based outliers in large datasets. In: Proceedings of the VLDB 1998, pp. 392–403 (1998)

[17] Knorr, E., Ng, R.: Finding intentional knowledge of distance-based outliers. In: Proceedings of the VLDB 1999, pp. 211–222 (1999)

[18] Jansen-Vullers, M.H., van der Aalst, W.M.P., Rosemann, M.: Mining configurable enterprise information systems. Data & Knowledge Engineering 56, 195–244 (2006)

[19] Agrawal, R., Gunopulos, D., Leymann, F.: Mining Process Models from Workflow Logs. In: Sixth International Conference on Extending Database Technology, pp. 469–483 (1998)

[20] Smith, H., Fingar, P.: Business Process Management: The Third Wave. Meghan-Kiffer Press, Tampa (2002)

[21] Yamanishi, K., Takeuchi, J.: A unifying framework for detecting outliers and change points from non-stationary time series data. In: KDD 2002, pp. 676–681 (2002)

[22] Yamanishi, K., Takeuchi, J.: Discovering outlier filtering rules from unlabeled data-combining a supervised learner with an unsupervised learner. In: Proceedings of the KDD 2001, pp. 389–394 (2001)

[23] Yamanishi, K., Takeuchi, J., Williams, G., Milne, P.: On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms. Data Mining and Knowledge Discovery, 275–300 (2004)

[24] He, Z., Xu, X., Huang, J.Z., Deng, S.: Mining class outliers: concepts, algorithms and applications in CRM

A Framework to Describe and Search for Virtual Resource Objects

Éric Renault, Wajdi Louati, Ines Houidi, and Houssem Medhioub

Institut Télécom – Télécom SudParis
Samovar UMR INT-CNRS 5157
firstname.lastname@it-sudparis.eu

Abstract. This paper presents a preliminary work on the development of a metalist-based description framework to efficiently describe and search for virtualized network resources for virtual network provisioning. Virtual resources are perceived as objects that can be described using metadata and semantically searched according to a set of criteria. The virtual resource objects are described by Infrastructure Providers that should advertise and offer their virtual resource descriptions to virtual network providers. These virtual network providers discover and match user queries for VNs with the virtual resources offered by the infrastructure providers. This work introduces the concept of metalist which is a simple and efficient way to specify sets of metadata for objects and explores how it is used to efficiently support virtual resource description.

1 Introduction

Network Virtualization has recently been seen as an enabler to design the Future Internet architecture. The virtualization concept ensures the coexistence of heterogeneous network architectures on a shared substrate allowing a diversified Internet and enabling the deployment of both evolutionary and revolutionary networking designs [1,2,3,4,5,6,7,8,9,10]. Network Virtualization also enables the emergence of new business roles (like infrastructure provider, VN provider/broker, etc.) [9,10,11,12] to offer on-demand virtual networks supporting customized services for end users.

One key principle in Network Virtualization is the description, discovery and matching of virtual resources offered by infrastructure (or resource) providers. Multiple specifications and languages have been proposed in the literature to describe network resources such as NDL [13], NM/perfSONAR [14], cNIS [15], VXDL [16], NML [17]. Unfortunately, these existing specifications do not provide schemes for describing virtual resources at the levels required for efficient resource searching, finding and matching. In previous work [18], a new virtual resource description schema has been proposed to enable infrastructure providers to describe and customize their offered virtual resources. Conceptual clustering techniques and similarity based matching have been also proposed in [18] to

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 208–219, 2010.

facilitate finding and matching of virtual resources. However, most of the description frameworks are still missing some basic functionalities needed at least in the scope of virtualization:

– many node and link descriptions are sharing lots of characteristics (this is one of the objectives of virtualization to provide the same of characteristics on different platforms).
– none of the description frameworks available so far are able to include dynamic information. For example, in the scope of network links, some metadata like the available user bandwidth or the one-way latency are changing over time according to the network traffic.
– not all descriptions shall be available to all users. Typically, an Infrastructure Provider probably would limit the access to its physical substrate to a subset of VN providers. As a result, some security and/or access control mechanism should be included.

As most of the applications for the Future Internet will rely on the ability to provide accurate information, a rich and flexible object description framework has become the key element for an efficient search. Many different models to enable semantic search have been developed over the past years, and the main important ones are Dublin Core [29], EXIF [23], IIM [22], OWL [25], RDF [24] and XMP [19]. Even though they all aim at organizing meta-information related to objects, there may be lots of differences between them. For example, some framework are one-domain oriented like Dublin Core for content and intellectual properties and EXIF for digital pictures description, while others like OWL, RDF and XMP allow users to define their own scheme.

To meet the resource description requirements cited above and in order to benefit from the rich and flexible descriptions offered by the meta-information principle, a virtual resource can be considered as an object that can be described using our proposed metalist model. The objective of this paper is to explore how the developed metalist model (further introduced in next sections) facilitates resource description, searching, finding and matching.

The paper is organized as follows. section 2 provides an overview of Network Virtualization and Network of information principles from the 4WARD project perspective [10]. Provisioning of virtual networks including description and matching of virtual resources is presented in Sec. 3. section 4 presents the metalist model and the benefits for the virtualization of networks. The prototype status of our proposal is given in Sec. 5.

2 The FP7 IP 4WARD Project

4WARD (Architecture and Design for the Future Internet) [10] is a European project that aims at overcoming the actual Internet impasses through: (1) enabling the co-existence of multiple networks on common platforms using virtualization of networking resources, (2) enhancing the utility of networks by making them self-managing, (3) increasing network robustness and efficiency by leveraging diversity, (4) improving application support by a new information-centric paradigm in place of the old host-centric approach.

The 4WARD project is structured into six activities:

- Business Innovation, Regulation, and Dissemination.
- New Architecture Principles and Concepts.
- Network Virtualization.
- In-Network Management.
- A new path abstraction (Generic Path).
- Networking of Information.

The work presented in this article only focuses on the two activities Network Virtualization and Networking of Information principles and explores how Virtual Network Provisioning can benefit from some Information-centric paradigms and models including the Metalist Model.

One of the basic concerns of 4WARD is to use the Network Virtualization concept to allow the coexistence of multiple networking solutions and architectures in the Future Internet. Virtualization provides a general approach to decouple the infrastructure from the services and to allow multiple network service providers to share a common physical infrastructure. Three main areas are considered in 4WARD:

1. Virtualization of both wired and wireless network resources.
2. Provisioning of large-scale on-demand Virtual Networks including the discovery of available physical and virtual resources, as well as the scalable embedding, control, and aggregation of these resources.
3. Management of instantiated virtual networks including control and dynamic re-allocation of resources during the lifetime of the virtual network.

Another main basic concern of 4WARD is moving the host-centric approach of the current Internet to an information-centric approach, ie. users should not worry any longer about how to access information, but should focus on the description of the objects they are looking for. This can be made possible by providing a set of metadata to published objects, and the Network of Information part of 4WARD aims at re-architect the access to information to enable this.

3 Virtual Network Provisioning

Network Virtualization provides separate logical networking environments, called *Virtual Networks*, over a shared networking infrastructure. Each Virtual Network (VN) appears to users as a dedicated physical network with dedicated network resources. Virtual networks can be supplied as a service by VN Provider(s) and delivered to VN users to support their infrastructures and services. There are essentially three main actors involved in the VN service delivery:

- the *Infrastructure Provider* is the actor that owns and controls the substrate network. The infrastructure providers ensure virtualization of their physical resources to offer virtual resources (as a service) to VN providers.

- the *VN Provider* allocates and aggregates virtual resources made available by Infrastructure Providers to establish VNs for users on demand.
- the *VN User* formulates and sends VN requests to VN Providers. The VN user corresponds to service providers, VN operators and end users.

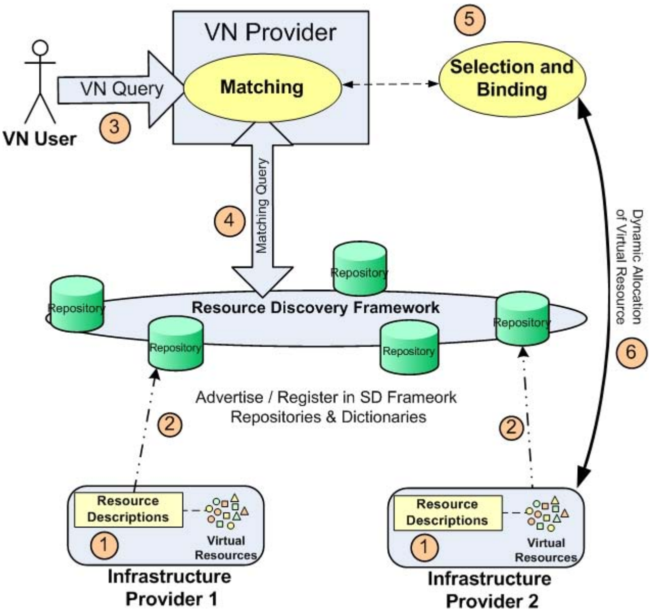


Fig. 1. Virtual Network Provisioning

The VN provisioning and setup consists in allocating a set of virtual resources extracted from substrate resources (e.g. substrate nodes, substrate links/paths) and assigning them to the required VN nodes and links specified in a VN request. As depicted in Fig. 1, five steps have been defined for the VN provisioning including description and advertisement (done by Infrastructure Providers), matching (typically achieved by VN providers), selection and binding (typically performed by Infrastructure Providers):

- Resource description: Infrastructure Providers have to describe the virtual resources and services they offer (step 1) to inform VN Providers about the virtual-resource basic properties but also about the degree of flexibility, availability and security of the offered resources. Virtual resources need to be described in terms of properties and functionalities much like nodes, devices and links are described in existing networking architectures.
- Resource advertisement: Infrastructure Providers advertise and register virtual resource descriptions in the discovery framework (step 2) using an

advertisement process (e.g. a publish/subscribe mechanism). VN providers use this framework to discover and match available resources using VN requests.
- Resource matching: upon receiving a VN request from users (step 3), resource matching consists in searching and finding a set of VN candidates (i.e. the appropriate virtual nodes and links) that fulfil the VN request. VN providers discover and match user queries for VNs with the virtual resources offered by Infrastructure Providers (step 4).
- Resource selection: among the matched resources, the selection step consists in choosing the best VN candidate using optimization algorithms (step 5).
- Resource binding allocates virtual resources from the substrate to set up the selected VN candidate (step 6).

4 The Metalist Model

The Network of Information, as defined in the 4WARD project, aims at managing meta-information about objects so as to perform more efficient semantic searches. Different kinds of objects have been defined [20], as shown on Fig. 2:

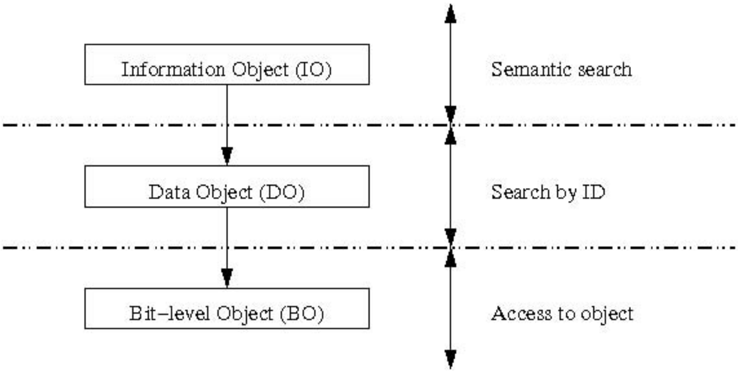


Fig. 2. The object model

- at the bottom are the *Bit-level Objects* (or BOs) that are storing the digital representation of objects, ie. the effective data for digital objects or any process that can communicate with the objects for non digital objects. In the scope of the virtualization of nodes and networks, this includes the virtual machines and the virtual network that interconnect them.
- on top of bit-level objects are *Data Objects* (or DOs) that are storing meta-information used by NetInf to manage the BOs. This includes location information to retrieve the BOs, security information to determine which users have the ability to access the object, etc.

- at the top, *Information Objects* (or IOs) are storing meta-information managed and used by users. There can be of any type and only users' use limits these meta-information.

From outside NetInf, ie. from users, services and/or applications point of view, only IOs and DOs are accessible, and they are the only ones to be associated an ID. DOs are not accessible, nor they are associated an ID. In fact, the nature of IOs and DOs is to store meta-information about the objects, while BOs are the objects themselves or a mean to access these objects. When a request is received by NetInf, it can be of two kinds: 1) a semantic search, ie. a set of conditions is provided and NetInf returns the list of IOs that matches the set; 2) a search by ID, ie. either a DO ID or an IO ID is received. If the ID belongs to an IO, the IO is returned. However, if the ID belongs to a DO, the object associated to the DO is returned to the user, not the DO itself (in other words, this means that if the object is digital, the user is sent back the object, while if the object is not digital, the user is directly connected to the digital representation of the object).

To represent the meta-information stored in both IOs and DOs, we defined the concept of metalist [27]. A metalist (tag NetInf:metalist) is a set of metadata that can be specified using three different ways (see Fig. 3):

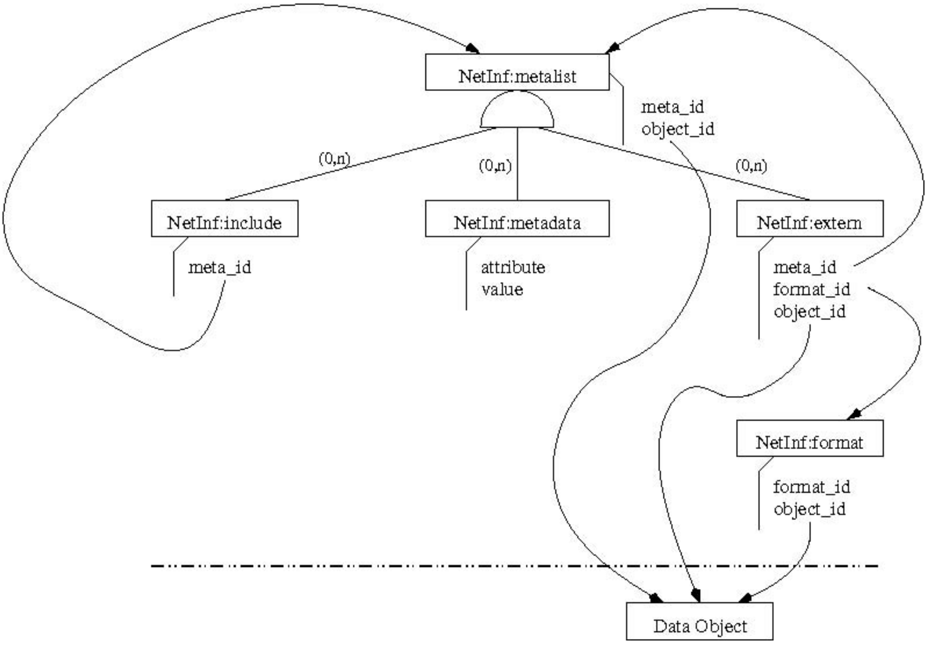


Fig. 3. The metalist model

- the metadata can be provided as is in the metalist (tag NetInf:metadata). A metadata is composed of two elements which are the value of the metadata and the attribute that tags the value if needed.
- the metadata can be imported from another metalist (tag NetInf:include). This allows to ensure consistency between the description of several objects if they share some meta-information.
- the metadata can be imported from another format type (tag NetInf:extern). This possibility allows to include external metadata that may be already existing in the set of metadata. For example, most digital cameras are including meta-information in digital pictures using the EXIF format [21]. The NetInf:extern mechanism avoids to copy these pre-existing metadata and ensures that metadata stored in NetInf are consistent with those available outside NetInf.

The metalist model offers the possibility to create classes to check the conformity of the metadata provided by the users according to a set of predicate, ie. if a set of mandatory attributes have been provided and/or if the value provided for these attributes are consistent. This is especially useful for the metadata stored in the DOs as they are directly used within NetInf. Two classes have been identified for use in the DO. The first one is used to locate the BOs that compose the DO. BOs may be replicated, cached or moved from one location to another and the corresponding DO must store this information to return BOs to the user. The second one is related to security and access control to object.

Both objects and their descriptions have to be protected so that only accurate information remains on the Future Internet. This security can be introduced using an access control to IOs and DOs. Access control can be of different types. It can be limited to the Unix *Read-Write-Execute* triplet or include more complex access control types like those used in databases (for example, the ability for a user to grant other users some specific access rights). Furthermore, other access rights that do not already exist shall be included in the future. Typical examples are those related to the copyright [26]. Some users (especially commercial companies) may be interested in providing public contents on the Future Internet making sure these contents will not be included inside uncontrolled documents by undesirable users. This appears to be a governance issue of the Future Internet and if applicable, these access rights will have to be managed. The metalist model allows to solve the problem generically as presented in [28].

5 Metalist-Based Virtual Resource Description Framework: Prototype Status

This section provides a UML diagram representing a virtual resource description schema (see Fig. 4). As shown in this diagram, a Network Element, seen as a basic building component, can be any of the following: Node, Link, Interface, or

Path. Each Network Element has an identifier (or name), an availability parameter, and functional as well as non-functional attributes. Functional attributes define characteristics, properties and functions of the network element (such as node/link type, execution environment, virtualization tool, OS, network stack, etc). Non-Functional attributes specify criteria and constraints related to the network element including performance, capacity, location, cost/price, QoS, etc.

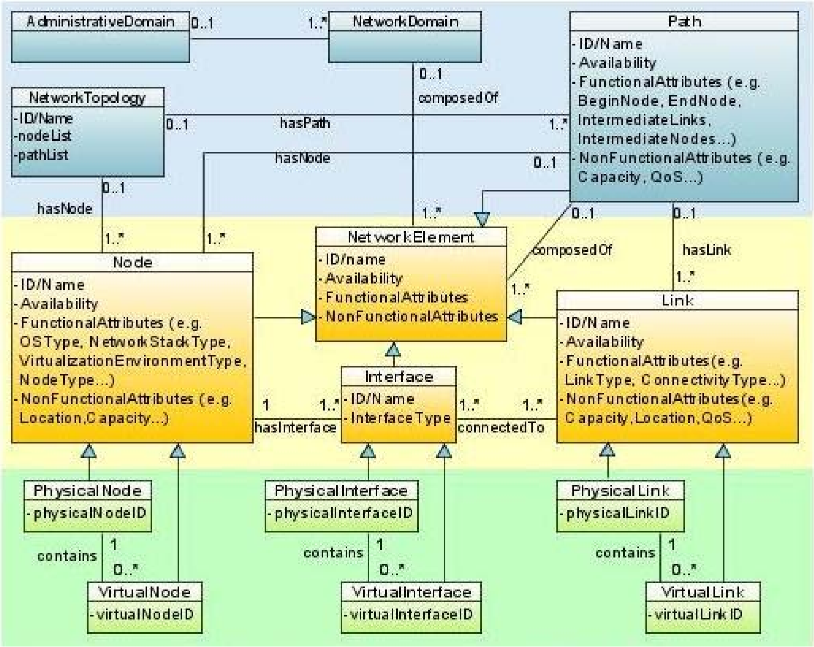


Fig. 4. The proposed virtual resource description schema

Each node in Fig. 4 has two sub-classes: virtual node and physical node. A physical node may contain one or multiple virtual nodes (e.g. virtual routers, virtual switches, virtual base stations). Each physical/virtual node has one or multiple physical/virtual interfaces. The interface types include, for example, Ethernet Interface, ATM Interface, Radio Interface (e.g. WiFi, WiMax) and Optical-Fiber Interface (SDH, SONET). One or multiple physical/virtual interfaces may be connected to one or multiple physical/virtual links. A physical link may support one or multiple virtual links (by using layer 1 and/or layer 2 virtualization). Each link has two additional characteristics: connectivity type (broadcast, point-to-point, point-to-multipoint) and QoS related parameters (such as bandwidth, delay, jitter, packet loss).

As depicted at the top of Fig. 4, the proposed schema defines an abstract class called Network Domain. It represents a group of Network Elements forming a specific network domain (e.g. an Autonomous System). Each Network Domain is managed by one Administrative Domain (e.g. infrastructure provider, VN provider). A Path is a group of Network Elements which can be composed of a single link or a sequence of links interconnected via intermediate nodes. The composition of a path (i.e. the set of intermediate nodes and links) is completely transparent to the VN provider which only sees the path endpoints. The NetworkTopology class represents the topology, the graph of connected network elements including all virtual and physical nodes and paths. Here we make no assumptions on the visibility of the physical resources given to the virtual network providers. Infrastructure providers actually decide how much information is disclosed to other parties on a case by case and SLA basis.

In this work, the extensible and standard XML schema has been used to express the proposed virtual resource description specification. This specification provides a common description framework for all infrastructure providers and ensures interoperability between their resource descriptions. Each node description includes the descriptions of all links directly connected to its network interfaces for completeness and simplicity reasons.

The main advantages of using the metalist model of the Network of Information for the description of virtual resources are the high flexibility of metadata management and the ability to offer users and applications a secure access and management. The prototype being developed for the Network of Information integrates two components: the first one is a storage space which aims at storing the digital representation of objects; the second one is an indexing space which aims at organizing the metadata provided by users, services and applications to tag objects. In the scope of virtualization, virtual resources are not effectively stored in the storage space and the Network of Information is mainly used to index virtual resource descriptions to allow efficient search.

For the representation of metadata, the metalist model has been implemented using the XML language and the eXist-db [30] database has been used for indexing. The storage space has been built on top of BitTorrent [31]. The storage space is intrinsically distributed over the network while the indexing space is centralized at present. A distributed version of the indexing space is scheduled to enhance redundancy and perform more efficient search. However, for the preliminary tests, the centralized implementation is acceptable.

In the scope of virtual resources, the description schema we developed has been adapted to the metalist model using the XML schema. An example of metalist is provided on Fig. 5 for both NetworkElement and Interface objects as depicted on Fig. 4. As a Link is also a NetworkElement, it may be clearer to separate the description of both parts into two different metalists, this resulting in the inclusion of the metadata associated to NetworkElement-1 in the set of metadata of Link-1.

```
<metalist meta_id="NetworkElement-1">
    <metadata attribute="ID">VL1</metadata>
    <metadata attribute="Name">Virtual Link #1</metadata>
    <metadata attribute="Availability">12/12/2009</metadata>
</metalist>

<metalist meta_id="Link-1" object_id="352CBF436DEF3245FABC56FD5E6">
    <metadata attribute="FA/LinkType">802.11</metadata>
    <metadata attribute="FA/ConnectivityType">Broadcast</metadata>
    <metadata attribute="NFA/Capacity">100</metadata>
    <metadata attribute="NFA/Location/Name">Site #1</metadata>
    <metadata attribute="NFA/Location/Address">Paris</metadata>
    <metadata attribute="NFA/QoS/Bandwidth">100 kbps</metadata>
    <metadata attribute="NFA/QoS/Latency">100 ms</metadata>
    <metadata attribute="NFA/QoS/Jitter">10 ms</metadata>
    <metadata attribute="NFA/QoS/Loss">2 %</metadata>
    <include meta_id="NetworkElement-1" />
</metalist>
```

Fig. 5. An example using the metalists

6 Conclusion

This paper addressed two main aspects of the 4WARD project, ie. the virtualization of network resources and the Network of Information. The paper highlighted how these two principles are being merged into a single consistent system allowing virtual resources, seen as objects, to be described and searched using the metalist model.

At present, both the virtualization of network resources and the metalist model have been implemented and are available as a prototype. We are currently merging these two prototypes into a single one. The next steps consist in developing and evaluating the performance of the proposed prototype and in emphasizing the impact of using the metalist model on top of eXist-db compared to the existing description and the search frameworks.

Acknowledgement

This work has been supported by the IST 7th Framework Programme Integrated Project 4WARD, which is partially funded by the Commission of the European Union. The views expressed in this paper are solely those of the authors and do not necessarily represent the views of Institut Télécom — Télécom SudParis or the respective projects and sponsors.

References

1. Global Environment for Network Innovations (GENI), http://www.geni.net
2. Anderson, T., Peterson, L., Shenker, S., Turner, J.: Overcoming the Internet impasse through virtualization. IEEE Computer Magazine 38(4), 34–41 (2005)
3. Turner, J., Taylor, D.E.: Diversifying the Internet. In: Proceedings of the IEEE Globecom 2005, St. Louis, MO, USA (November 2005)
4. Niebert, N., et al.: The Way 4WARD to the Creation of a Future Internet. In: Proc. of the 19th International Symposium on Personal, Indoor and Mobile Radio Communications. Cannes, France (September 2008)
5. Bavier, A., Feamster, N., Huang, M., Peterson, L., Rexford, J.: In VINI veritas: Realistic and controlled network experimentation. In: Proceedings of ACM SIGCOMM, Pisa, Italy (September 2006)
6. Touch, J., et al.: A Global X-Bone for Network Experiments. In: Proceedings of the IEEE Tridentcom 2005, Trento, Italy, pp. 194–203 (March 2005)
7. He, J., et al.: DaVinci: Dynamically Adaptive Virtual Networks for a Customized Internet. In: Proceedings of the CoNex 2008 (December 2008)
8. Bhatia, S., et al.: Hosting virtual networks on commodity hardware. Georgia Tech Computer Science TR-GT-CS-07-10 (January 2008)
9. Feamster, N., Gao, L., Rexford, J.: How to lease the Internet in your spare time. SIGCOMM Comput. Commun. Rev. 37(1), 61–64 (2007)
10. The FP7 4WARD Project, http://www.4ward-project.eu
11. Louati, W.: On demand Virtual Network Service for Dynamic Networks. PH.D Thesis Number 07INT003, UPMC & GET-INT, FRANCE. (March 2007), http://www-public.int-edu.eu/~louati/Publications/thesis.pdf
12. Zhu, Y., Zhang-Shen, R., Rangarajan, S., Rexford, J.: Cabernet:between Connectivity architecture for better network services. In: Proceedings of the Workshop on Rearchitecting the Internet (December 2008)
13. NDL web page, http://www.science.uva.nl/research/sne/ndl/
14. OGF Network Measurement Working Group, http://nmwg.internet2.edu/
15. Wolski, M., Osinski, S., Gruszczynski, P., Labedzki, M., Patil, A., Thomson, I.: Deliverable DS3.13.1: common Network Information Service Schema Specification. GEANT2 (2007)
16. Koslovski, G., Vicat-Blanc Primet, P., Charão, A.S.: VXDL: Virtual Resources and Interconnection Networks DescriptionLanguage. GridNets (2008)
17. NML-WG – Network Mark-up Language Working Group, http://www.ogf.org/gf/group_info/view.php?group=nml-wg
18. Houidi, I., Louati, W., Zeghlache, D., Baucke, S.: Virtual Resource Description and Clustering for Virtual Network Discovery. In: ICC Workshop on the Network of the Future 2009 (2009)
19. XMP Specification. Adobe Systems Incorporated (September 2005)
20. Dannewitz, C., Pentikousis, K., Rembarz, R., Renault, É., Strandberg, O., Ubillos, J.: Scenarios and Research Issues for a Network of Information. In: MobiMedia 2008, Oulu, Finland (July 2008)
21. Exchangeable Image File Format for Digital Still Cameras: Exif Version 2.2. Standard of Japan Electronics and Information Technology Industries Association (April 2002)
22. IPTC-NAA Information Interchange Model, version 4.1. International Press and Telecommunications Council (July 1999)

23. Exchangeable Image File Format for Digital Still Cameras: Exif Version 2.2. Standard of Japan Electronics and Information Technology Industries Association (April 2002)
24. Manola, F., Miller, E.: RDF Primer. W3C Recommendation (February 2004)
25. Dean, M., Schreiber, G.: OWL Web Ontology Language Reference. W3C Recommendation (February 2004)
26. Rao, S.: Copyright: its implications for electronic information. Online information review 27(4), 264–275 (2003)
27. Renault, É., Zeghlache, D.: The Metalist Model: a Simple and Extensible Information Model for the Future Internet. In: Eunice 2009 – The Internet of the Future, Barcelona, Spain (September 2009)
28. Renault, É., Ahmad, A., Abid, M.: Toward a Security Model for the Future Network of Information. In: International Workshop on Ubiquitous Computing & Applications, Fukuoka, Japan (December 2009) (to appear)
29. Weibel, S., Kunze, J., Lagoze, C., Wolf, M.: Dublin Core Metadata for Resource Discovery. In: RFC 2413 (September 1998)
30. Meier, W.: eXist: An Open Source Native XML Database. In: Chaudhri, A.B., Jeckle, M., Rahm, E., Unland, R. (eds.) NODe-WS 2002. LNCS, vol. 2593, pp. 169–183. Springer, Heidelberg (2003)
31. Bit torrent web page, http://www.bittorrent.com/

Developer Toolkit for Embedded Fuzzy System Based on E-Fuzz

C. Chantrapornchai, K. Sripanomwan, O. Chaowalit, and J. Pipatpaisarn

Department of Computing, Faculty of Science,
Silpakorn University, Thailand 73000
ctana@su.ac.th

Abstract. In this work, we propose a development toolkit, called E-Fuzz-Wizard to help fuzzy system designers for designing embedded fuzzy systems. The toolkit composes of software and hardware that enables creating the rapid prototype. It contains the examples which use the hardware and code generated to produce a prototype. The software has a visual interface which allows the user to specify the requirement of fuzzy systems in terms of the fuzzy set characteristics, inference methods, rules and defuzzification method. It generates the code in C that is runable in the chosen microcontroller platform. E-Fuzz Wizard also integrates unique features such as concurrent and real-time fuzzy system design as well as hardware mapping and customization. The generated code will facilitate the embedded fuzzy system development process. The toolkit is easy to use and facilitate the beginners to develop a fuzzy system.

Keywords: Embedded Systems, E-Fuzz, Fuzzy Design Tools.

1 Introduction

In Thailand, fuzzy system development is really required knowledge from experts in the field such as fuzzy controls. Also, to learn about the fuzzy systems, it is difficult for Thai students and teachers in high school to understand the use of fuzzy logic in every day life. According to the education policy by the government, it is urged to develop embedded system experts in the country. Many secondary schools and universities participate in the competitions related to embedded system fields in any platform. Fuzzy logic is a means to control many embedded equipments. It would be necessary to promote the fuzzy knowledge in such a field so that the domain experts will be expanded. It is found that even the teachers in high schools do not know fuzzy logic and its benefits. It is also difficult for them to understand in a short period. To help them understand better to gain more practical knowledge, it would be benefit if we have a laboratory fuzzy toolkit for the students to learn fuzzy logic development and its practices in the embedded world.

Fuzzy systems are now being used in many consumer electronic devices such as air conditions, washing machines, refrigerators etc. To implement a fuzzy system, one may choose to implement using many means such as general-purposed processors, fuzzy processors, programmable devices etc. In consumer electronics, the cost of the

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 220–233, 2010.

device is important. Microcontrollers are the good choice since they are convenient to find and not expensive. However, the capability of the microcontrollers is very limited. Further programming on microcontrollers needs skills at the low level. Thus, together with complication in fuzzy systems itself, it is not very easy and convenient to create such systems on microcontrollers.

To develop a fuzzy control system, one needs to go through many steps to find the right parameters. These parameters are membership functions, rules, defuzzification methods, and inference methods. A lot of software simulations need to be done to select the proper parameters. Each set of parameters yield distinct system characteristics, e.g., static and dynamic memory usage, execution speed, and accuracy. After the parameters are selected, the fuzzy system is then implemented on the particular hardware[4].

The objective of this research is to develop a toolkit to help building the rapid prototype of simple embedded fuzzy applications. Also, the toolkit should facilitate the study of fuzzy system development in the laboratory which shows the practice of embedded fuzzy systems. The toolkit contains the hardware, software, and laboratory manual. The software is called E-Fuzz Wizard. It has a visual interface in a drag and drop manner. The user specifies the fuzzy system requirement, i.e., the membership function shape, the inference methods, rules, and defuzzification method. Then the code in C for the specified platform is generated. The tool also includes the simulator which enables the testing of the fuzzy requirements before real implementation. Besides, the tool supports the concurrent and real-time features which allow the more subsystems to be run and communicate concurrently. It allows the user to specify the mapping on the physical device to view the testing results when applying the hardware. The tool then generates C fuzzy application prototypes on microcontrollers as well as FPGA. The hardware sample and the laboratory manual guide the sample development of typical prototype fuzzy systems.

Several works have been done in developing fuzzy control hardware. Most fuzzy processors often have limitation such as 2 inputs and 1 output rules [7,9,22], the shape of the membership is triangular or trapezoid. Some work is based on analogue systems [10,15,17]. Some are specific to applications e.g. pattern recognition [20]. Some requires extra hardware supports and special instruction sets [19]. Many works are based on VLSI systems such as [16,21]. Ascia and Catania [3] presented a VLSI hardware for fuzzy processing. The hardware can handle 8 inputs, 4 outputs and 256 rules. Chen et.al. [5] proposed a high speed parameterized fuzzy processor. The processor considers parallel and pipeline processing. A general-purposed CPU is another choice which is flexible but may be too much powerful. Microcontrollers often provide a moderate solution since it is programmable and easy to create a prototype. However, many computations in fuzzy systems are expensive, such as the use of floating point, multiplication and division etc. They are not suitable to the 8-bit microcontroller unless certain optimization is done.

Many previous works have mentioned about the fuzzy development software and tools. Nishidai and Hajimi [12] present a tool which consists of hardware for fuzzy rule reasoning. Nishiuchi and Masamitsu [3] presented an approach to generate code for a fuzzy control program. Ahmed et.al. [1] presented an adaptive fuzzy software management tool. The tool can cope with fuzziness in the software development process. Iqbal et.al. [2] proposed a fuzzy expert system for a manufacturing process, particularly in a machining process. The expert system can adapt and learn automatically. Zhang and Kandel [23] proposed a CPU scheduling method using fuzzy logic under a multiple criteria. Mateou and Andreou [11] presents a framework to develop decision support systems which uses fuzzy cognitive maps and genetic algorithms. Rasmussen and Yager [14] developed a fuzzy query language called SummarySQL which is able to perform a smart query and search for data mining. One of the tools that are close to ours is fuzzyTECH. It is the commercial work of fuzzyTECH which provides fuzzy libraries. Their target platform is based on MCS51 and MCS96[25]. Xfuzzy is another well-known software tools which enable the developments of fuzzy[29]. It provides the multiple fuzzy system definitions and generates the code in C/VHDL. Though both works are applicable, it does not have an explicit support for the concurrent features.

Using our tool, it is convenient to understand the fuzzy system development. The visual software aids the system design. The users only put the whole system architecture and define each system's characteristics. Then the simulation helps view the behavior of the systems. Once the design is settled, the code implementation is generated. It can be programmed in the given hardware. Once hardware is programmed, the system is ready to use.

Our tool has special characteristics:
In the software, we overcome the limitation of existing ones such as the fixed number of inputs and the limited shape form of the membership function. It provides a flexible interface which allows designers to create many system inputs, the shapes can be varied, and many of multiple-input rules may be used. Using the tool, developers can learn to also create a more complex system by adding the concurrent real-time features. Such features are unique and not commonly found in any fuzzy expert system design tool. The code generated is in C on PIC microcontrollers with the good estimation of memory usages and timing characteristics.

The hardware included are PIC microcontroller board whose CPU can be replaced. The designer can use their own PIC CPU or the given one in the set.

The code generated is programmed by any existing IDE such as MPLAB or NIOS II. Thus, the hardware platform can be flexible. The given hardware in the set is for convenience in developing systems the student laboratory purpose.

The paper is organized as following: next section presents fundamental fuzzy system components and the inference process and related computation. Section 3 presents the structure of E-Fuzz Wizard and its capabilities. Section 4 presents an example of UI interface and development methods using the tool. Section 5 shows the application examples. Section 6 concludes the paper and discusses the future work.

2 Backgrounds

Typical fuzzy systems have the characteristics as shown in Figure 1.

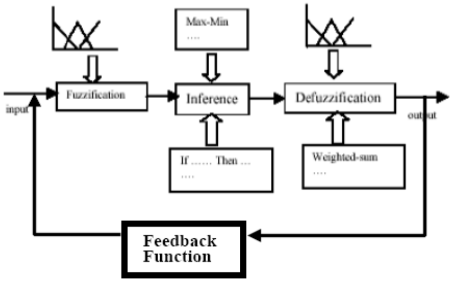


Fig. 1. Fuzzy System Components

In Figure 1, a given input is read and then fuzzified to be a fuzzy set. This is called the fuzzification process. After that, the value is given to the inference engine to find out which rules are fired. For each rule that is fired, the corresponding output linguistic variable is marked. This step is called a fuzzy inference process. Then, all the fired output linguistic values are concluded to be a crisp value which is the actual output. This is called the defuzzification process. The output is given to the feedback function which is a computation of some linear/nonlinear function and it becomes the input again.

Based on these steps, necessary parameters that one needs to come up with when the system is designed are the input and output variables, linguistic variables for each input and output and their membership functions , fuzzy rules, inference method and defuzzification method.

In the following, we briefly explain steps to establish a fuzzy system according to the above components [18].

1. Inputs and outputs: First designers need to define the number of inputs and output. The universes of each input and output are defined. This defines domain for each fuzzy set.

2. Linguistic variables and membership functions: For each input and output, one needs to define the set of linguistic variables. Each linguistic variable corresponds to a membership function which specifies a mapping from an element to a degree of membership value ranged [0,1]. The membership function is typically defined by triangular shape, trapezoidal shape, bell shape, etc.

3. Rule set: From the given inputs and output and linguistic variables for each one, the set of rules are defined. Typically, the rules are defined from all possible combinations of input linguistic variables. After that, the rule optimization may be done to minimize the number of rules, the number of inputs for each rule, and to minimize the memory usage for the rule set. Most fuzzy systems require 2 inputs and one output. Many fuzzy hardware implements the 2-dimensional associative memory for the rule set.

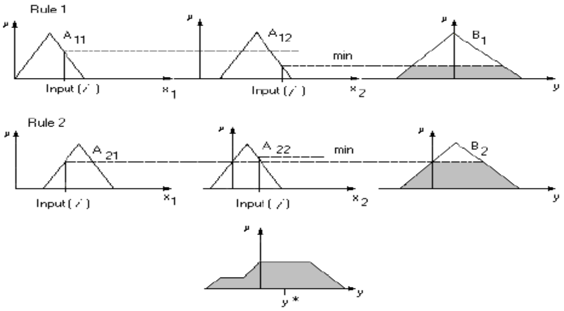


Fig. 2. Max-min inference

4. Inference method: Designers need to specify the fuzzy inference method used to compute the output membership degree. Many operators are defined in the literature [23]. The commonly used one is max-min or max-product. Figure 2 shows a graphical example of using max-min inference[18]. The minimum value between two membership values of the two inputs are used as the cut to the output linguistic variable for each rule. Then the fuzzy set for each output linguistic variable is unioned to become a final set and the final set is defuzzified. For max-product, rather using min operation, the product operation is used instead.

$$z^* = \frac{\int \mu_{\underset{\sim}{C}}(z) \cdot z dz}{\int \mu_{\underset{\sim}{C}}(z) dz} \tag{1}$$

5. Defuzzification method: From Figure 2, the final set is defuzzified to get a crisp output value y*. Several methods can be used to defuzzify such as using the max value (means of max, smallest of max,largest of max), computing a centriod value, using the approaches such as weighted-average, center of sum. The commonly used one is centroid which is computed as the equation (1) and is depicted by Figure 3.

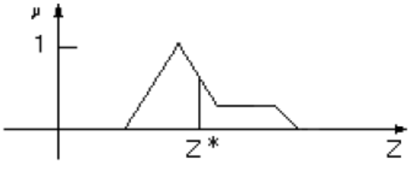


Fig. 3. Centriod defuzzification

Another example of the simpler defuzzification method is weighted-average. This is described by the equation (2) and Figure 4.

$$z^* = \frac{\sum \mu_{\underset{\sim}{C}}(\bar{z}) \cdot \bar{z}}{\sum \mu_{\underset{\sim}{C}}(\bar{z})} \tag{2}$$

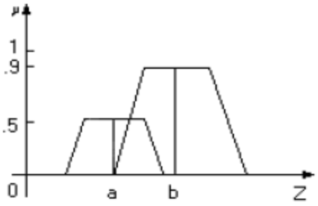


Fig. 4. Weighed-average defuzzification

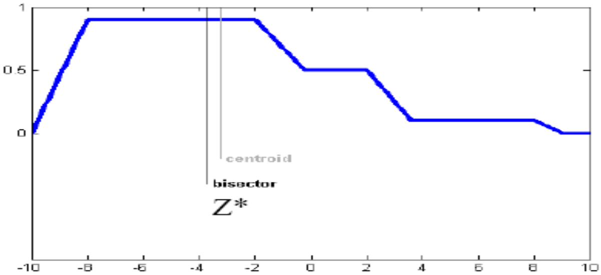
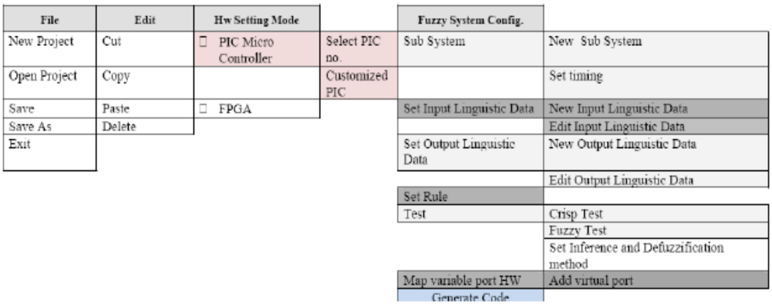


Fig. 5. Centroid VS Bisector

Another example for defuzzification method which has a good performance is bisector approach. The approach computes the vertical line z* which divides into two equal regions. Figure 5 compare the points obtained by centriod and bisector [28].

3 E-Fuzz Software and Components

In E-Fuzz, the target architectures are FPGA and PIC microcontrollers. The current version of the software focuses on the microcontroller. Figure 6 shows the E-Fuzz menu for a user to develop a particular fuzzy system.



File	Edit	Hw Setting Mode		Fuzzy System Config.	
New Project	Cut	□ PIC Micro Controller	Select PIC no.	Sub System	New Sub System
Open Project	Copy		Customized PIC		Set timing
Save	Paste	□ FPGA		Set Input Linguistic Data	New Input Linguistic Data
Save As	Delete				Edit Input Linguistic Data
Exit				Set Output Linguistic Data	New Output Linguistic Data
					Edit Output Linguistic Data
				Set Rule	
				Test	Crisp Test
					Fuzzy Test
					Set Inference and Defuzzification method
				Map variable port HW	Add virtual port
				Generate Code	

Fig. 6. E-Fuzz Menu

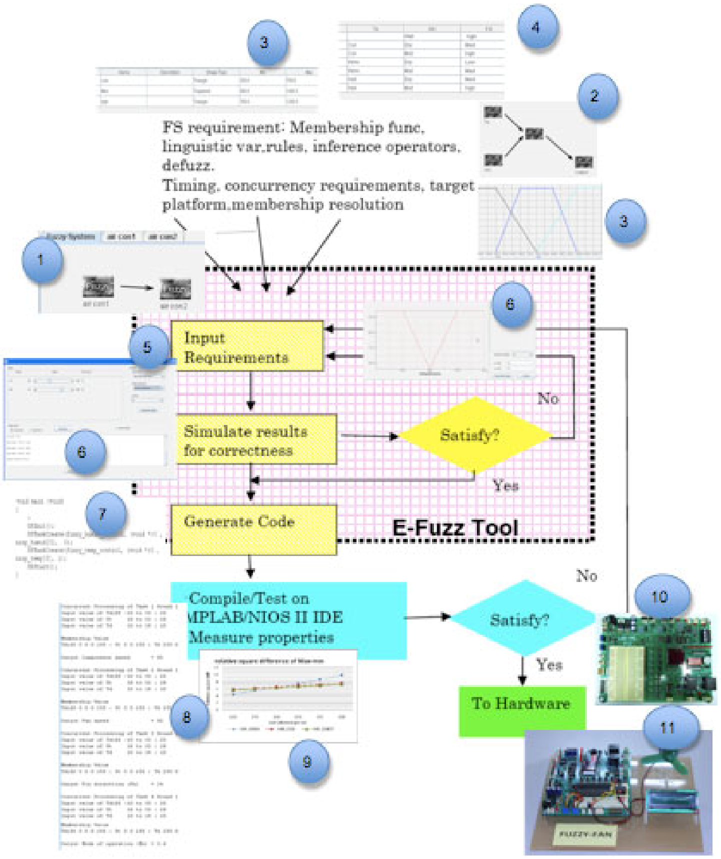


Fig. 7. Usage of E-Fuzz Tool

From E-Fuzz Components in Figure 6, a user needs to specify the overall architecture of the system first. He specifies the details of fuzzy systems and their interconnections. Each fuzzy system is defined based on the given parameters such as input/output domain, linguistic variables, inference method, defuzzification method, etc. Each system can be simulated by its own using either crisp or fuzzy input test against the given feedback function. The code for each fuzzy system can then be generated in the specified platform after the user satisfies with the parameter setting. For the overall system, the communications between fuzzy systems can be given as output-input relation. As the hardware is specified, the port mapping of input/output of fuzzy systems to the real device port or virtual port can be done. Virtual ports are defined by the operating system level such as mailbox and queue in Micro C/OS II. Each fuzzy system may be run periodically. Inputs may also be read periodically. Once these timing properties are given and the period is specified, the main code containing periodic fuzzy systems and function codes are generated.

Figure 7 shows the framework of using the software. The developer defines the system architecture. (1) displays the whole system. (2) displays each system architecture.

Then each system parameter are defined: (3) inputs, output, linguistic variables, (4) rules, (5) inference method and defuzzification. After that, he verifies the behavior of the system in the tool (6). Once the behavior is satisfied, the code implementation can be generated automatically (7). Then the code is programmed using the existing tools (8). The code properties may be verified again in these tools (9). Since C code is in a high and each IDE has a specific complier, the code low-level property depends on it. The developer may need to tune the code to adjust in details. Then the code is programmed into the board in the toolkit (10) and accessories are connected for practice (11). In the testing, we use MPLAB (www.microchip.com) or NIOS II (www.altera.com) since we are interested in developing the prototype in PIC and FPGA. Each of these components phase will be described in the next sections.

4 Case Study Dialogs

Figure 8 shows the case where we design a system containing two connected subsystem. On the edge, we can specify a way to communicate the output as an input on another system.

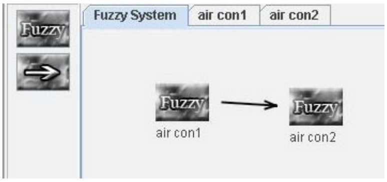


Fig. 8. Main Diagram

Then the user specifies each fuzzy system component's properties. This is as shown in Figure 9. In Figure 9, we develop a fuzzy temperature control. The system needs two inputs which are humidity and temperature values and gives one output which is a fan speed setting.

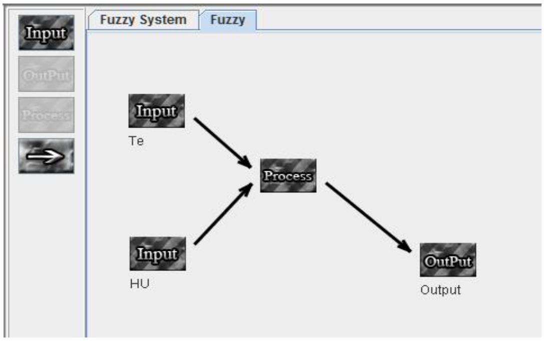
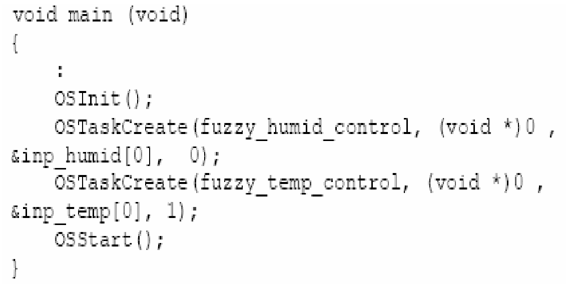


Fig. 9. Specification of fuzzy components



```
void main (void)
{
    :
    OSInit();
    OSTaskCreate(fuzzy_humid_control, (void *)0 ,
&inp_humid[0],  0);
    OSTaskCreate(fuzzy_temp_control, (void *)0 ,
&inp_temp[0], 1);
    OSStart();
}
```

Fig. 10. Main code for concurrent fuzzy systems

After all the specification is done, the code is generated. Figure 10 shows the pseudo code of the generated code for concurrent systems. The code to generate is pretty much straightforward.

```
void fuzzy_humid_control (void *pdata) {

        while (1) {

                Fuzzy_humid();

                OSTimeDlyHMSM(0,0,1,0);// delay 1 second

        }

}
```

Fig. 11. Fuzzy Task with delay

For this case we develop two fuzzy systems. We use Micro C/OS II to support the concurrency fuzzy systems. Each system is defined as a task in Micro C/OS II. Each fuzzy task is implemented in a loop with some delay as a period in Figure 11. It is shown that each fuzzy system in the diagram is mapped to a function corresponding to a task. The communication between fuzzy tasks is done by mail box or queue in Micro C/OS II. Using the library, the code functions necessary for fuzzy computation are reused.

5 Hardware Kit and Examples

In the toolkit, we also provide a set of hardware and its peripherals to help set up the fuzzy logic laboratory. The hardware board is based on PIC CPU. The board has the palette (420 holes) to stick the PIC processor on. It contains dip switches, 2 DC inputs which are suitable for connecting sensors, relay circuit, LED, serial port, reset circuit, RTC, 7-segment display, buzzer, D/A for testing typical examples in fuzzy controls. The LAB-PIC board is shown in Figure 12. The example of PIC that comes with the kit is shown in Figure 13. This one is PIC18F6627.

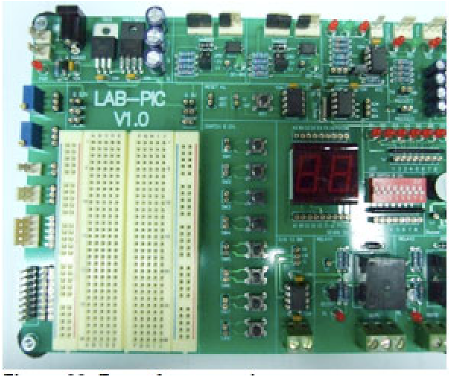


Fig. 12. LAB-PIC Board

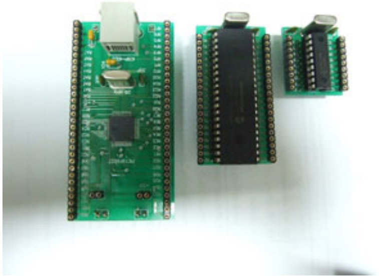


Fig. 13. PIC chip example that is in the kit (PIC18F6627)

We also include the manuals which present easy steps for fuzzy design examples using the software to help learning fuzzy system developments. Based on the fuzzy fan example shown in the previous section, we simplify it and take the generated code and program into the LAB-PIC. We connect the input to the resistors assuming they are the sensor reading of both temperature and humidity (Figure 14). We also connect the output control to control the speed of the fan. The output is also displayed on the LCD shown in the figure. Figure 15 shows the LAB-PIC and hardware connection for testing.

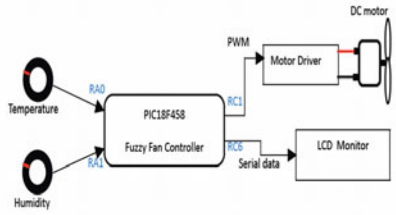


Fig. 14. Fuzzy fan structure

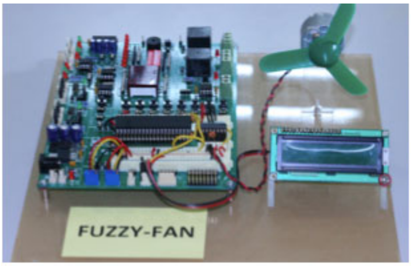


Fig. 15. Fuzzy fan example on LAB-PIC

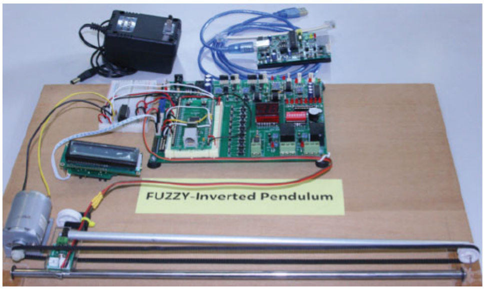


Fig. 16. Fuzzy Inverted Pendulum example on LAB-PIC

We also have another example on an inverted pendulum [18]. In this example, we are to control the aluminum ruler. The input of the system is the angle read by the sensor (assuming it is the resistant value). The output of the system is the PWM value to control the direction to move the motor and to control the speed of the motor. The motor controls the standing of the aluminum ruler. The setup kit for this laboratory is shown in Figure 16.

Table 1. Comparison between E-Fuzz and others

Aspects	XFuzzy	FuzzTech5.7	E-Fuzz
Fuzzification- Membership function specification and Membership function code	-Specify graph parameters. Software draws the graphs. Support both curve and linear functions. -Generate as a function computing the value	-Drag graph for both curve, linear functions. N/A *	-Drag graph for both curve, linear functions. -Store points in array. Generate a function to map a value.

Table 1. (*continued*)

Rules Specification and Code	-Input as table, matrix, rule formats -Generate a code function with inference operator	-Input as rule, matrix formats N/A*	-Input as a rule format -Generate as code together with inference operators.
Defuzz.	CoG,MoM,LoM,FoM, Takogi, Weighted fuzzy mean, Gamma,Max	CoM,MoM	CoG,MOM,Bisect, LoM,Weight-average
Multiple systems/Concurrency	Yes Code generated individually for each step of fuzzy process as each function per system.	Yes N/A*	Yes Code generated can be run as tasks under Micro OS/II Tasks can be communicated using virtual using Mailbox/Queue or physical port of MCU.
Timing	No	No	-there is a timing associated with input readings and each system.
Language	C,C++,Java,VHDL	C ,Java,C++,VB,Matlab	C
Target platform	Not Specified	MCU,PLC	PIC, FPGA
Simulation/Debugging	Yes	Yes	Yes
Training support	Yes	Yes	No
Commercial	No	Yes	No

6 Conclusion

In this work, we present the toolkit for developing fuzzy systems. The kit contains 1) hardware 2) visual tool called E-Fuzz Wizard. 3) Sample laboratories. The hardware is designed containing standard interface and I/O used by simple control applications. The software provides the integrated facility to build the concurrent real-time and embedded fuzzy systems. It provides ways to specify each fuzzy system parameter visually. Real-time properties of the fuzzy systems can be specified. The software includes the simulation of various parameter setting of the fuzzy systems. Once the user satisfies with the parameter selection, the code for the target platform is generated. The current version targets at platforms : microcontroller PIC and FPGA. The generated code is written in C for microcontrollers with embedded real-time OS (micro C/OS II). The code is programmed to the hardware using existing programming IDE tools. The examples show a real practice of the designed fuzzy systems. All together, the kit gives an easy way to build the prototype fuzzy system for the beginners: from parameter selection, behavior tuning, code generation, until hardware mapping.

Acknowledgement

This work is supported in part by NECTEC, NSTDA, Thailand grant number NT-B-22-ES-17-51-01 and Faculty of Science Funding, Silpakorn University, Thailand.

References

[1] Ahmed, M.A., Saliu, M.O., AlGhamdi, J.: Adaptive Fuzzy Logic-Based Framework For Software Development Effort Prediction. Information and Software Technology 47, 31–48 (2005)

[2] Iqbal, A., Khan, I., Dar, N.U., He, N.: A Self-Developing Fuzzy Expert System, Designed for Optimization of Machining Process. In: Proceedings of the World Congress on Engineering, vol. III (2008)

[3] Ascia, G., Catania, V.: An Efficient Hardware Architecture to Support Complex Fuzzy Reasoning. International Journal on Artificial Intelligence Tools 5(1-2), 41–60 (1996)

[4] Chantrapornchai, C.: Rapid prototyping Methodology and Environment for Fuzzy Applications. In: Optimization Techniques 1973. LNCS, vol. 4, pp. 940–949. Springer, Heidelberg (2003)

[5] Chen, B.T., Chen, Y.S., Hsu, W.H.: Performance evaluation of a parameterized fuzzy processor (PFP). Fuzzy sets and systems 81(3), 293–309 (1996)

[6] Frías-Martínez, E.: Design of a Lukasiewicz rule-driven fuzzy processor. Soft Computing - A Fusion of Foundations, Methodologies and Applications 7(1), 65–71 (2002)

[7] Gabrielli, E.G., Masetti, M.: Design of a family of VLSI high speed fuzzy processors. In: IEEE Fuzz 1996 (1996)

[8] Ghaus, C.: Fuzzy model and control of a fan-coil. Energy and Buildings Journal 33, 545–551 (2001)

[9] Falchieri, D., Gabrielli, A., Gandolfi, E.: Very fast rate 2-input fuzzy processor for high energy physics. Fuzzy Sets and Systems 132, 261–272 (2002)

[10] Li, J.H., Lim, M.H., Cao, Q.: Evolvable Fuzzy Hardware for Real-time Embedded Control in Packet Switching. Evolvable Machines 161, 205–227 (2005)

[11] Mateou, N.H., Andreou, A.S.: A framework for developing intelligent decision support systems using evolutionary fuzzy cognitive maps. Journal of Intelligent and Fuzzy Systems 19(27), 151–170 (2008)

[12] Nishidai, Hajimi: Fuzzy reasoning and methods, rule setting apparatus and methods. Eurpoean Patent Classification (1997): G06F 9/44. Publication number: EP0513829, `http://www.freepatentsonline.com/EP0513829.html`

[13] Fumitaka, N., Masamitsu, I.: Method for generating fuzzy control program. Japanese Patent no. JP7160306.3 (1995), `http://www.sumobrain.com/patents/jp/Method-generating-fuzzy-control-program/JP07160306.html`

[14] Rasmussen, D., Yager, R.R.: SummarySQL - A Fuzzy Tool For Data Mining. Intelligent Data Analysis (1997)

[15] Song, C.T.P., Quigley, S.F., Pammu, S.: Novel analogue fuzzy inference processor. In: Proceedings of ISCAS, vol. 3, pp. 247–250 (1998)

[16] Pagni, A., et al.: Automatic Synthesis Analysis Implementation of a Fuzzy Controller. In: IEEE Int'l Conf. Fuzzy Systems, pp. 105–110. IEEE Process, Piscataway (1993)

[17] Pammu, S.: Novel Analogue Fuzzy Inference Processor. In: Proceedings of ISCAS, vol. 3, pp. 247–250 (1998)

[18] Ross, T.J.: Fuzzy Sets. In: Fuzzy Logic and Fuzzy Systems: Theory and Applications, McGraw Hill, New York (1995)

[19] Salpura, V., Gschwind, M.: Hardware/Software Co-Design of a Fuzzy RISC Processor. Proceedings of the IEEE 83, 422–434 (1995)

[20] Shi, B., Lin, G.: Programmable and expandable fuzzy processor for pattern recognition. United States Patent 6272476 (2001), http://d.wanfangdata.com.cn/Periodical_dianzixb200002008.aspx

[21] Masaki, T., Hiroyuki, W.: A VLSI implementation of a fuzzy inference engine: toward an expert system on a chip. International Journal on Information Sciences 38, 147–163 (1986)

[22] Tsutomu, M.: Fuzzy processor, European Patent EP0392494 (1990)

[23] Viot Greg, J., Sibigtrogth James, M., Broseghinl James, L.: A Method for performing a fuzzy logic operation in data processor. European Patent: EP0574714 (2000)

[24] Zhang, Y.-Q., Kandel, A.: Fuzzy CPU Scheduling. International Journal on Artificial Intelligence Tools 6(2), 211–225 (1997)

[25] http://www.fuzzytech.com

[26] http://www.cs.cmu.edu/afs/cs/project/ai-respository/ai/areas/fuzzy/systems/fuzzyfan

[27] http://www.mathworks.de/products/demos/shipping/fuzzy/defuzzdm.html#3

[28] http://www.micrium.com

[29] http://www.imse.cnm.es/Xfuzzy

[30] http://www.programmersheaven.com/download/1244/download.aspx

A Quadsection Algorithm for Grammar-Based Image Compression

Morihiro Hayashida, Peiying Ruan, and Tatsuya Akutsu

Bioinformatics Center, Institute for Chemical Research, Kyoto University,
Uji, Kyoto 611-0011, Japan
{morihiro,ruan,takutsu}@kuicr.kyoto-u.ac.jp

Abstract. Grammar-based compression is to find a small grammar that generates a given data and has been well-studied in text compression. In this paper, we apply this methodology to compression of rectangular image data. We first define a context-free rectangular image grammar (CFRIG) by extending the context-free grammar. Then we propose a quadsection type algorithm by extending a bisection type algorithm for grammar-based compression of text data. We show that our proposed algorithm approximates in polynomial time the smallest CFRIG within a factor of $O(n^{4/3})$, where an input image data is of size $O(n) \times O(n)$. We also present results on computational experiments on the proposed algorithm.

Keywords: Bisection, Context-free Rectangular Image Grammar.

1 Introduction

Image compression is one of well-studied problems in data compression and image processing. Extensive studies have been done on image compression, and several methods and/or formats such as JPEG, GIF, PNG have been widely used.

Various techniques are employed in these widely used methods. JPEG was named after Joint Photographic Experts Group, and is usually lossy compression for photographic still images. Each block of size 8×8 pixels is transformed using two-dimensional DCT (Discrete Cosine Transform). The higher frequency components are more coarsely reduced by quantization. Finally, the image is compressed using Huffman coding [4]. GIF stands for Graphics Interchange Format, and is lossless compression for images with less than or equal to 256 distinct colors, based on the Lempel-Ziv algorithm [10], which is a dictionary coder that reads a sequence, constructs a dictionary dynamically, and replaces the sequence with words of the dictionary. PNG stands for Portable Network Graphics, and has been developed to replace GIF. PNG uses filtering and Deflate algorithm that is the combination of the Lempel-Ziv algorithm [10] based method and Huffman coding [4]. The compression rate of PNG is often higher than that of GIF.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 234–248, 2010.

Though compression ratios of these methods are very high for most image data, there are cases where some of these methods fail to achieve high compression ratios. Furthermore, in many existing methods, compressed data are difficult to interpret. That is, it is difficult to extract some patterns, which exist in the original image, from compressed data.

On the other hand, in text compression, extensive studies have been done on grammar-based compression [2,6,8], which is to find a small grammar generating a given string. It is useful not only for data compression but also for extraction of repetitive patterns. Therefore, it is reasonable to try to study grammar-based compression for image data. Various grammars have been proposed for producing image data [3,9]. However, to our knowledge, there was no grammar-based image compression algorithm with a guaranteed approximation ratio. Therefore, in this paper, we extend grammar-based compression for text data to image data compression. In particular, we present QUADSECTION algorithm that is obtained by extending BISECTION algorithm for text data compression [2,6]. Furthermore, we show that QUADSECTION computes in polynomial time a grammar of size $O(g^* n^{4/3})$ for a given image of size $O(n) \times O(n)$, where g^* is the size of a minimum grammar generating the given image.

The organization of the paper is as follows. In Section 2, we define a context-free rectangular image grammar by extending the context-free grammar, formalize the smallest grammar problem for image data, and prove the NP-hardness of the problem. Next, we present QUADSECTION in Section 3, analyze its approximation ratio in Section 4, and extend it to higher dimensions in Section 5. Then, we present results on some computational experiments in Section 6. Finally, we conclude with future work.

2 Context-Free Rectangular Image Grammar

Here, we define *CFRIG* (Context-Free Rectangular Image Grammar). A CFRIG is defined by a 4-tuple $(\Sigma, \Gamma, S, \Delta)$ where Σ, Γ, $S \in \Gamma$ and Δ are a set of terminal symbols, a set of nonterminal symbols, the start symbol and a set of production rules, respectively. Each terminal symbol corresponds to a label of a pixel, and is denoted by a lower-case letter. Each nonterminal symbol corresponds to a rectangular region, and is denoted by an upper-case letter. Since each nonterminal symbol is associated with a rectangular region, each nonterminal symbol is represented as $A_{n,m}$, which means that this symbol generates an image with $n \times m$ pixels (i.e., an image composed of n rows and m columns). Then, we consider the following two types of production rules

(R1) $A_{1,1} \rightarrow a$,
(R2) $A_{n,m} \rightarrow [B_{n_1,m_1}, C_{n_1,m_2}; D_{n_2,m_1}, E_{n_2,m_2}]$,
where $n_1 + n_2 = n$ and $m_1 + m_2 = m$.

The meanings of these rules are clear from Fig. 1. For a rule of type (R2), we allow subcase (R2') of $n_2 = 0$ (i.e., D_{n_2,m_1} and E_{n_2,m_2} are empty) and subcase (R2") of $m_2 = 0$ (i.e., C_{n_1,m_2} and E_{n_2,m_2} are empty). We write $A_{n,m} \rightarrow [B_{n,m_1}, C_{n,m_2}]$

and $A_{n,m} \rightarrow [B_{n_1,m}; D_{n_2,m}]$ for the former case and latter case, respectively (see also Fig. 1 (R2') and (R2")).

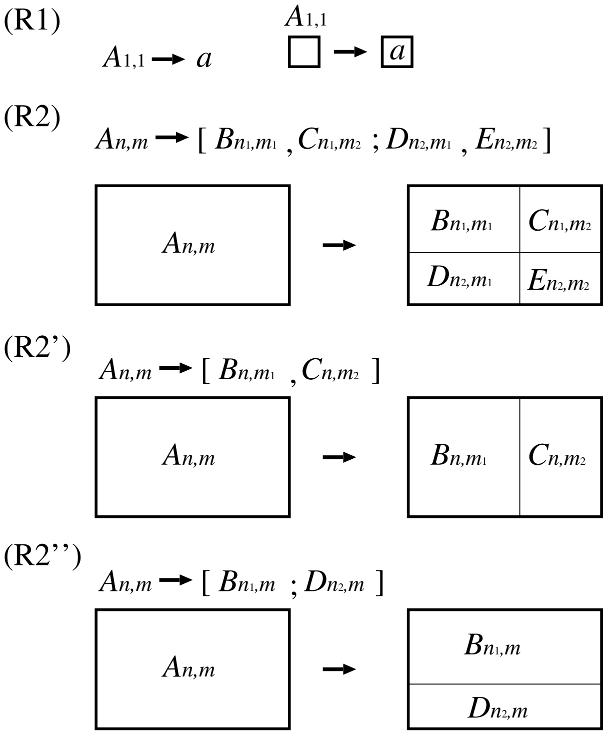


Fig. 1. Production rules for CFRIG

The *size of a grammar* is defined as the total number of symbols appearing in the right hand sides (RHSs) of production rules. From the definition of CFRIG, it is seen that only acyclic grammars are allowed in CFRIG. Furthermore, when we discuss compression algorithms, as in [2], we only consider *non-ambiguous* CFRIGs, that is, each nonterminal symbol appears in the left hand side (LHS) of exactly one rule.

Though we restricted the form of rules to (R1) and (R2), more general rules can be represented by using multiple rules of type (R1) and (R2), as shown in Fig. 2. We can show that such a transformation increases the size of the grammar only by a constant factor though we omit the proof here.

Based on the above definitions, we define the smallest grammar problem for image data is to find a smallest CFRIG which uniquely generates a given image of size $n \times m$. We can show that the smallest grammar problem for image data is NP-hard.

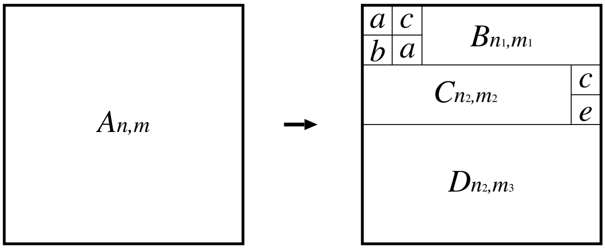


Fig. 2. This kind of rules can be transformed into CFRIG with a constant factor increase of the size

Theorem 1. *Finding the smallest CFRIG for a given image data is NP-hard.*

Proof. Since a string of length m is regarded as an image of size $1 \times m$, we can use almost the same reduction as in [2]. However, since we can only use rules of types of (R1) and (R2), we need to slightly modify the reduction.

Let $G(V, E)$ be an instance of the vertex cover problem. Let $N = |V|$ and $M = |E|$. Recall that the vertex cover problem asks whether or not there exists $W \subseteq V$ of size k such that for any edge $\{u, v\} \in E$, $u \in W$ or $v \in W$ holds.

From G and k, we construct an instance of CFRIG as follows. We map G to an image I of size $1 \times m$ by

$$I = \prod_{v_i \in V} (\#v_i|v_i\#|)^2 \prod_{\{v_i,v_j\} \in E} (\#v_i\#v_j\#|),$$

where v_i denotes a distinct terminal corresponding to each vertex, each '|' denotes a distinct terminal (delimiter), and xy means a concatenation of x and y. Let W be a vertex cover of size k. Then, we will have the following rules.

$$\begin{aligned} D_j &\to |_j, \\ H &\to \#, \\ A_i^0 &\to v_i, \\ A_i^L &\to A_i^0 H, \\ A_i^R &\to H A_i^0, \\ A_i &\to A_i^R H, \qquad \text{if } v_i \in W \end{aligned}$$

where $|_j$'s are introduced since each '|' denotes a distinct delimiter. It is straightforward to verify that the total size of these production rules is

$$4N + M + 1 + N + 2N + 2N + 2k = 9N + M + 2k + 1.$$

If a production rule with long RHS were allowed, we would have such a rule as

$$\begin{aligned} S \to\ & A_1^R D_1 A_1^L D_2 A_1^R D_3 A_1^L D_4 \cdots \\ & A_1 A_2^L D_{4N+1} A_1 A_3^L D_{4N+2} \cdots, \end{aligned}$$

for the start symbol S. It is to be noted that for each edge $\{v_i, v_j\} \in E$, a subsequence '$A_i A_j^L D_{4N+l}$' or '$A_i^R A_j D_{4N+l}$' appears in this rule. If $v_i \in W$, the former appears. Otherwise, the latter appears. It is straight-forward to see that the size of this rule is

$$8N + 3M.$$

In order to represent this rule by CFRIG, we need

$$8N + 3M - 1$$

rules of type (R2'), where each rule is of size 2. Summing up all, the total size of a grammar corresponding to W is

$$25N + 7M + 2k - 1.$$

Therefore, there exists a grammar of size $25N + 7M + 2k - 1$ that generates image I if there exists a vertex cover of size k.

On the other hand, suppose that there exists a grammar of size at most $25N + 7M + 2k - 1$ which generates image I. Then, as in the proof of Theorem 1 in [2], we need only consider grammars having the above mentioned form. Then, we can construct a vertex cover of size k from the set of nonterminals each of which has an expansion of the form $\#v_i\#$. Therefore, the theorem holds. □

We can prove that CFRIG remains NP-hard even if $n \times n$ images are given, where the details are omitted in this paper.

3 Compression Algorithm

Our compression algorithm for image data is based on BISECTION [2,6] and is denoted by QUADSECTION here. BISECTION takes a string, and recursively decomposes the string into two smaller substrings. QUADSECTION recursively decomposes a given rectangular image $I_{n,m}$ into smaller rectangular images until each image consists of one pixel, where the same nonterminal symbol is assigned to identical rectangular images. Let $h(i)$ be 2^j for the largest integer j such that $2^j < i$, where we let $h(1) = 1$. For example, $h(2^i) = 2^{i-1}, h(2^i + 1) = 2^i$. For a rectangular image $I_{n,m}$ of size $n \times m$, $I_{[i_1:i_2],[j_1:j_2]}$ denotes the sub-rectangular image composed of i_1th - i_2th rows and j_1th - j_2th columns. The following is a pseudocode of QUADSECTION, where it is invoked with the input image $I_{n,m}$ and an empty grammar $\mathcal{G}$. QUADSECTION returns the start symbol that generates $I_{n,m}$.

```
procedure QUADSECTION($I_{n,m}$)
  var
    $I_{n,m}$: an image of size $n \times m$;
    $A_{n,m}$: a nonterminal symbol uniquely assigned to $I_{n,m}$;
    $n, m, h, n_1, n_2, m_1, m_2$: Integer;
```

```
begin
   if the same image $I'_{n,m}$ as $I_{n,m}$ has already appeared then return $A'_{n,m}$;
   if $n = 1$ and $m = 1$ then
      add $A_{1,1} \to a$ to $\mathcal{G}$ where $I_{n,m} = a$;
      return $A_{1,1}$;
   endif;
   $h := \max\{h(n), h(m)\}$;  $n_1 := \min\{n, h\}$;  $m_1 := \min\{m, h\}$;
   $n_2 := n - n_1$;  $m_2 := m - m_1$;
   if $n_1 = n$ then
      $B_{n,m_1} :=$ QUADSECTION($I_{[1:n],[1:m_1]}$);
      $C_{n,m_2} :=$ QUADSECTION($I_{[1:n],[m_1+1:m]}$);
      add $A_{n,m} \to [B_{n,m_1}, C_{n,m_2}]$ to $\mathcal{G}$;
   else if $m_1 = m$ then
      $B_{n_1,m} :=$ QUADSECTION($I_{[1:n_1],[1:m]}$);
      $D_{n_2,m} :=$ QUADSECTION($I_{[n_1+1:n],[1:m]}$);
      add $A_{n,m} \to [B_{n_1,m}; D_{n_2,m}]$ to $\mathcal{G}$;
   else
      $B_{n_1,m_1} :=$ QUADSECTION($I_{[1:n_1],[1:m_1]}$);
      $C_{n_1,m_2} :=$ QUADSECTION($I_{[1:n_1],[m_1+1:m]}$);
      $D_{n_2,m_1} :=$ QUADSECTION($I_{[n_1+1:n],[1:m_1]}$);
      $E_{n_2,m_2} :=$ QUADSECTION($I_{[n_1+1:n],[m_1+1:m]}$);
      add $A_{n,m} \to [B_{n_1,m_1}, C_{n_1,m_2}; D_{n_2,m_1}, E_{n_2,m_2}]$ to $\mathcal{G}$;
   endif;
   return $A_{n,m}$;
end.
```

It is straight-forward to see that QUADSECTION works in polynomial time.

4 Analysis

In the following, we assume without loss of generality (w.l.o.g.) that $n \geq m$. If we consider images with $n \times 1$ pixels, CFRIG corresponds to CFG and thus the lower bounds on the approximation ratio on compression in [2] holds for CFRIG. In the same way, the lower bound for BISECTION (Theorem 5 in [2]) holds also for QUADSECTION.

Proposition 1. *The approximation ratio of* QUADSECTION *is* $\Omega(\sqrt{n}/\log n)$.

Similarly, we obtain the following proposition.

Proposition 2. *The smallest CFRIG that generates an image of size* $n \times m$ *has size* $\Omega(\log n)$.

In order to analyze the upper bound of QUADSECTION, we first establish mk Lemma [2] for CFRIG, where we use g instead of m to denote the size of a grammar here.

Lemma 1. *If the input image data $I_{n,m}$ is generated by an CFRIG of size g, $I_{n,m}$ contains at most $2ngk$ distinct sub-images of size $k \times h$, where we assume w.l.o.g. that $k \geq h$.*

Proof. Let $H_{k,h}$ be a sub-image of size $k \times h$ of $I_{n,m}$. If $k = h = 1$, $H_{1,1}$ is represented by RHS of a rule of type (R1). Otherwise, since CFRIG is an acyclic grammar, there exists a rule of type (R2), $A_{n',m'} \rightarrow [B_{n_1,m_1}, C_{n_1,m_2}; D_{n_2,m_1}, E_{n_2,m_2}]$, that $I_{A_{n',m'}}$ contains $H_{k,h}$ and none of $I_{B_{n_1,m_1}}$, $I_{C_{n_1,m_2}}$, $I_{D_{n_2,m_1}}$ and $I_{E_{n_2,m_2}}$ contains $H_{k,h}$, where I_A for a nonterminal A denotes the expansion image of A. (See Fig. 3.) We assume w.l.o.g that a part of $H_{k,h}$ is included in $I_{B_{n_1,m_1}}$. $H_{k,h}$ is one of at most $km_1 + hn_1 \leq 2nk$ sub-images. Therefore, $I_{n,m}$ contains at most $2ngk$ distinct sub-images of size $k \times h$. □

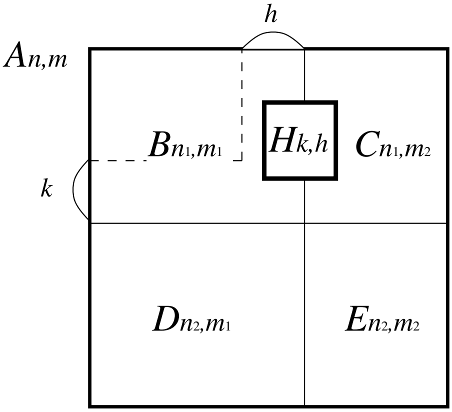


Fig. 3. Proof of Lemma 1

Theorem 2. QUADSECTION *computes in polynomial time an CFRIG of size $O(g^* n^{4/3})$ for a given image $I_{n,m}$ of size $n \times m$ ($n \geq m$), where g^* is the size of the smallest CFRIG generating $I_{n,m}$.*

Proof. We prove the theorem only for the case that $n = m = 2^l$ holds for some integer l. Modification of the proof for the other cases is straight-forward.

The number of sub-images that are generated by recursive calls of depth at most k is bounded by

$$1 + 4 + 4^2 + \cdots + 4^k.$$

On the other hand, the number of distinct sub-images that are generated by recursive calls of depth at least $k+1$ is bounded by

$$2 \sum_{i=0}^{h} g^* n 2^i$$

from Lemma 1, where $h = \log n - k$. Therefore, the number of production rules generated by QUADSECTION is

$$O(4^k + g^* n 2^{\log n - k}).$$

By letting $4^k = n2^{\log n - k}$, we have

$$k = \frac{2}{3} \log n.$$

Therefore, the number of production rules generated by QUADSECTION is

$$O(n^{4/3} + g^* n^{4/3}) = O(g^* n^{4/3}). \qquad \square$$

It is to be noted that the grammar may have size $O(n^2)$ in the worst case and thus the above approximation ratio is meaningful.

5 Extension to *d*-Dimensional Volume Data

Although we have defined a grammar for two-dimensional image data, we can define another grammar, *d-CFRVG* (Context-Free *d*-dimensional Rectangular Volume Grammar), for *d*-dimensional volume data ($d \geq 3$) in a similar way. A d-CFRVG is defined by a 4-tuple $(\Sigma, \Gamma, S, \Delta)$ where Σ, Γ, $S \in \Gamma$ and Δ are a set of terminal symbols, a set of nonterminal symbols, the start symbol and a set of production rules, respectively. Each terminal symbol corresponds to a label of a d-dimensional voxel, and is denoted by a lower-case letter. Each nonterminal symbol corresponds to a d-dimensional rectangular region, and is denoted by an upper-case letter. Since each nonterminal symbol is associated with a d-dimensional rectangular region, each nonterminal symbol is represented as $A_{n^{(1)},\cdots,n^{(d)}}$, which means that this symbol generates a volume with $n^{(1)} \times \cdots \times n^{(d)}$ units. Then, production rules are as follows.

(R1) $A_{\underbrace{1,\cdots,1}_{d}} \to a$,

(R2) $A_{n^{(1)},\cdots,n^{(d)}} \to \left[B_{n_1^{(1)},n_{j_2}^{(2)},\cdots,n_{j_d}^{(d)}}, C_{n_2^{(1)},n_{j_2}^{(2)},\cdots,n_{j_d}^{(d)}} \right]_{(j_2,\cdots,j_d)\in\{1,2\}^{d-1}}$,
where $n_1^{(i)} + n_2^{(i)} = n^{(i)}$ for $i = 1, \cdots, d$.

It should be noted that d-CFRVG is also an acyclic grammar. For d-CFRVG, the following lemma holds as well as CFRIG.

Lemma 2. *If the input volume data $V_{n^{(1)},\cdots,n^{(d)}}$,where $n = n^{(1)} \geq \cdots \geq n^{(d)}$, is generated by an d-CFRVG of size g, $V_{n^{(1)},\cdots,n^{(d)}}$ contains at most $dkgn^{d-1}$ distinct sub-images of size $k^{(1)} \times \cdots \times k^{(d)}$, where we assume w.l.o.g that $k = k^{(1)} \geq \cdots \geq k^{(d)}$.*

QUADSECTION can be extended in a straight-manner to compression of d-dimensional volumes, and is called HYPERSECTION.

Theorem 3. HYPERSECTION *computes in polynomial time a d-CFRVG of size* $O(g^* n^{d^2/(d+1)})$ *for a given volume* $V_{n^{(1)},\cdots,n^{(d)}}$ *of size* $n^{(1)} \times \cdots \times n^{(d)}$*, where* g^* *is the size of the smallest d-CFRVG generating* $V_{n^{(1)},\cdots,n^{(d)}}$*.*

Proof. We prove the theorem only for the case that $n = n^{(1)} = \cdots = n^{(d)} = 2^l$ holds for some integer l as well as Theorem 2.

The number of sub-volumes that are generated by recursive calls of depth at most k is bounded by

$$1 + 2^d + 2^{2d} + \cdots + 2^{kd}.$$

On the other hand,the number of distinct sub-volumes that are generated by recursive calls of depth at least $k+1$ is bounded by

$$d \sum_{i=0}^{h} g^* n^{d-1} 2^i$$

from Lemma 2, where $h = \log n - k$. Therefore, the number of production rules generated by HYPERSECTION is

$$O(2^{kd} + g^* n^{d-1} 2^{\log n - k}).$$

By letting $2^{kd} = n^{d-1} 2^{\log n - k}$, we have

$$k = \frac{d}{d+1} \log n.$$

Therefore, the number of production rules generated by HYPERSECTION is

$$O(n^{d^2/(d+1)} + g^* n^{d^2/(d+1)}) = O(g^* n^{d^2/(d+1)}).$$

□

It is to be noted that the grammar may have size $O(n^d)$ in the worst case and thus the above approximation ratio is meaningful.

6 Computational Experiments

We implemented QUADSECTION and applied it to several images. In our implementation, input raw images are given in PGM (Portable GrayMap) format or PPM (Portable PixMap) format. An image in PGM and PPM format consists of the format type, width, height, maximum pixel value, and pixel values in raster scan order. Each pixel value is represented either by ascii codes or by binary values according to the format type. Since the file size in ascii format depends on the pixel values, we used only binary format. A pixel value in PGM format is stored in 8 bits, and that in PPM format is stored in 24 bits, where each color of red, green and blue is represented in 8 bits.

The implemented version of QUADSECTION generates CFRIG $(\varSigma, \varGamma, S, \varDelta)$ from a given image in PGM or PPM format, and the grammar is stored in newly introduced QSN format as follows (see Table 1). The set of rules $\varDelta$ is divided into four sets of rules, $\varDelta^{(\mathrm{R2})}$, $\varDelta^{(\mathrm{R2'})}$, $\varDelta^{(\mathrm{R2''})}$ and $\varDelta^{(\mathrm{R1})}$, corresponding to the types of rules, (R2), (R2'), (R2") and (R1) respectively. Let $\varDelta_i$ denote the i-th rule of $\varDelta$. Then, $\varDelta_1 = \varDelta_1^{(\mathrm{R2})}$, $\varDelta_{|\varDelta^{(\mathrm{R2})}|+1} = \varDelta_1^{(\mathrm{R2'})}$ and $\varDelta_{|\varDelta|} = \varDelta^{(\mathrm{R1})}_{|\varDelta^{(\mathrm{R1})}|}$. In particular, we suppose that LHS of either $\varDelta_1^{(\mathrm{R2})}$, $\varDelta_1^{(\mathrm{R2'})}$, $\varDelta_1^{(\mathrm{R2''})}$ or $\varDelta_1^{(\mathrm{R1})}$ is the start symbol S. The nonterminal symbol of LHS of $\varDelta_i$ is replaced with the number $i-1$. Thus, each nonterminal symbol is represented with $\lceil \log |\varDelta| \rceil$ bits number. Each terminal symbol, that is a pixel value, is represented with 8 bits for PGM format and 24 bits for PPM format. In Table 1, RHS$(\varDelta_i)$ denotes the nonterminal and terminal symbols appeared in RHS of $\varDelta_i$. In QSN format, the numbers corresponding to symbols contained in RHS$(\varDelta_i)$ are stored sequentially in order. In the case of the black-color image of size 512 × 512, QUADSECTION generates the following 10 rules: $S = A_{512,512} \rightarrow [A_{256,256}, A_{256,256}; A_{256,256}, A_{256,256}]$, $\cdots$, $A_{2,2} \rightarrow [A_{1,1}, A_{1,1}; A_{1,1}, A_{1,1}]$, $A_{1,1} \rightarrow 0$. Since $\lceil \log |\varDelta| \rceil = \lceil \log 10 \rceil = 4$, $|\varDelta^{(\mathrm{R2})}| = 9$ and $|\varDelta^{(\mathrm{R1})}| = 1$, the compressed file size in QSN format is $8 + 1 + 8 + 2 + \lceil \log 10 \rceil \cdot (3 + 4 \cdot 9) + 8 \cdot 1 = 183$ bits. It should be noted that the actual file size is $\lceil 183/8 \rceil = 23$ bytes because files are created in a storage in terms of bytes.

Table 1. QSN format for CFRIG $(\varSigma, \varGamma, S, \varDelta)$

# bits	contents
8	maximum pixel value
1	0 if PGM, 1 if PPM
8	$\lvert\varDelta\rvert$
2	rule type including the start symbol
$\lceil \log \lvert\varDelta\rvert \rceil$	$\lvert\varDelta^{(\mathrm{R2})}\rvert$
$\lceil \log \lvert\varDelta\rvert \rceil$	$\lvert\varDelta^{(\mathrm{R2'})}\rvert$
$\lceil \log \lvert\varDelta\rvert \rceil$	$\lvert\varDelta^{(\mathrm{R2''})}\rvert$
$4\lceil \log \lvert\varDelta\rvert \rceil$ /rule	RHS$(\varDelta_i^{(\mathrm{R2})})$ $(i = 1, \cdots, \lvert\varDelta^{(\mathrm{R2})}\rvert)$
$2\lceil \log \lvert\varDelta\rvert \rceil$ /rule	RHS$(\varDelta_i^{(\mathrm{R2'})})$ $(i = 1, \cdots, \lvert\varDelta^{(\mathrm{R2'})}\rvert)$
$2\lceil \log \lvert\varDelta\rvert \rceil$ /rule	RHS$(\varDelta_i^{(\mathrm{R2''})})$ $(i = 1, \cdots, \lvert\varDelta^{(\mathrm{R2''})}\rvert)$
8 (or 24) /rule	RHS$(\varDelta_i^{(\mathrm{R1})})$ $(i = 1, \cdots, \lvert\varDelta^{(\mathrm{R1})}\rvert)$

In order to evaluate the compression ability of QUADSECTION, we compared the following image file formats, PNG (Portable Network Graphics), GIF (Graphics Interchange Format), JPEG (Joint Photographic Experts Group) and IFS (Iterated Function Systems) [7]. Both of GIF and PNG use lossless compression algorithms as well as QUADSECTION. It should be noted that GIF is not able to deal with more than 256 distinct colors. If an image has more than 256 distinct colors, GIF ignores less frequently used colors. In order to enhance compression rates, PNG firstly employs filtering, which replaces the color of each pixel with the difference of colors between adjacent pixels. It makes use of the characteristics that colors of adjacent pixels are often very close in images. IFS is a quadtree-based fractal image coder/decoder, and the software called Mars implemented by Polvere [7] is available from `http://inls.ucsd.edu/~fisher/Fractals/Mars-1.0.tar.gz`.

We examined the following images, 'black', 'cross' (Fig. 4), 'cross2' (Fig. 5), 'hilbert' (Fig. 6), 'lena' (Fig. 7), and 'lena2' (Fig. 8). The image of black is considered before, consists of black color pixels, and the size is 512 × 512. The image of cross consists of 3 distinct colors, and the size is 512 × 512. The image of cross2 is the left-top part of that of cross, and the size is 234 × 345. The image of hilbert is a 6-th order Hilbert curve. The Hilbert curve is known as one of fractal diagrams, can be formed using the following rules. $S \rightarrow A$, $A \rightarrow LBFRAFARFBL$, $B \rightarrow RAFLBFBLFAR$, where S is the start symbol, L means 'turn left at a right angle', R means 'turn right at a right angle', and F means 'draw forward' [3]. The image of lena was transformed from the full color image file '4.2.04.tiff' provided on the USC-SIPI Image Database (`http://sipi.usc.edu/database/`) to PGM format using a tool of ImageMagick (`http://www.imagemagick.org/`), 'convert', and the size is the same as the original one, 512 × 512. The image of lena2 was transformed from '4.2.04.tiff' to the binary image in PGM format using '-colors 2' option of the tool 'convert'. We also used the tool 'convert' in order to transform the above images from PGM format to PNG, GIF or JPEG format.

Table 2. Results on the compression sizes (byte) in QSN, PNG, GIF, JPEG and IFS formats for several images

image	QSN	PNG	GIF	JPEG	IFS
black	23	265	828	1185	1544
cross	57	1688	8645	2392	1544
cross2	146	1031	2797	885	1114
hilbert	181	1196	11083	99172	40582
lena	697387	223614	264340	65338	15825
lena2	33524	17859	15501	22888	2868

Table 2 shows the results on the compression sizes (byte) in QSN, PNG, GIF, JPEG, and IFS formats for the images. The uncompressed size of black, cross, hilbert, lena, and lena2, in PGM format is 262159 bytes, and that of cross2 is 80745 bytes, respectively. For the images having symmetric and geometric patterns, that is black, cross, and hilbert, QUADSECTION was able to compress them better than other image compression methods. However, for the image of cross2, the compression size in QSN format was larger than that of cross

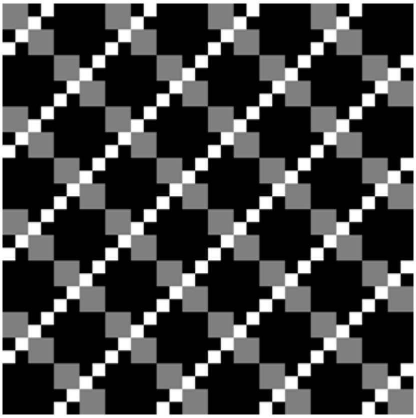


Fig. 4. Image of cross with size 512 × 512 and 3 distinct colors

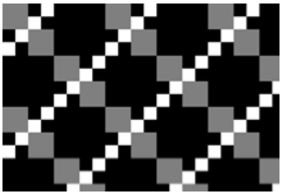


Fig. 5. Image of cross2 with size 234 × 345. cross2 is the left-top part of cross

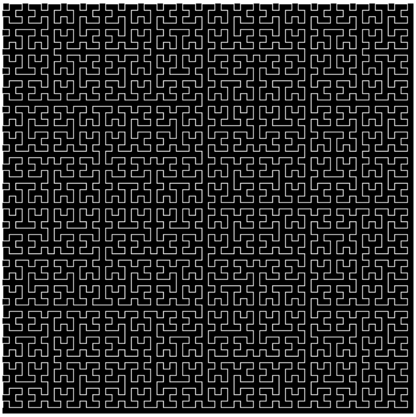


Fig. 6. Binary image of hilbert with size 512 × 512

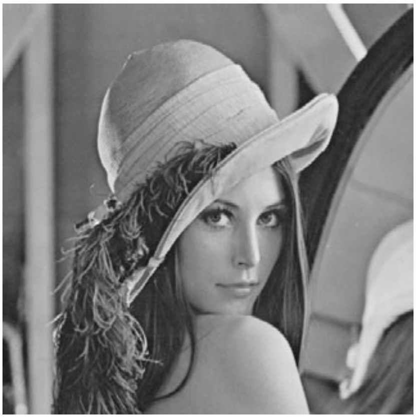


Fig. 7. Gray-scale image of lena with size 512 × 512

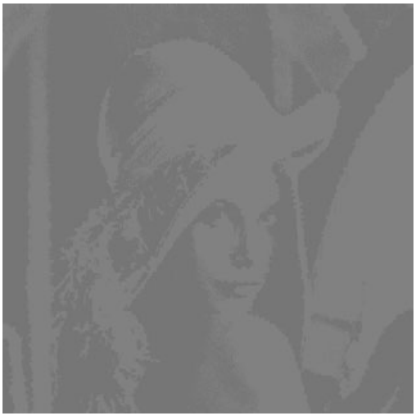


Fig. 8. Binary image of lena2 with size 512 × 512

although the image size of cross2 is smaller than cross, and the compression sizes of cross2 in other formats were smaller than those of cross. It is considered that the image size of cross2, 234 × 345, is not a power of two, some sub-images corresponding to nonterminal symbols had various sizes, and it increased the number of rules. It should be noted that IFS did not compress the hilbert image well. It is considered because fractal image compression methods find sub-images whose contraction image is similar to a sub-image. In contrast, the results of IFS for the photographic images, lena and lena2, were good. For the image of lena, which is a gray-scale photographic image, the compression size in QSN format was larger than the size of the raw image and the compression sizes in other image formats. It is considered that QUADSECTION could not compress it well because the image is not symmetric and has many colors. The compression size of lena2 in QSN format was smaller than that of lena and the size of lena2 in PGM format. The rate of the compression size in QSN format to that in PNG or GIF format decreased from about 3 for lena to about 2 for lena2. This result suggests that QUADSECTION is still useful for compression of non-symmetric binary image data.

7 Conclusion

We have proposed a grammar-based image compression algorithm, QUADSECTION, by extending the BISECTION algorithm for text data compression. For that purpose, we defined CFRIG, which is an extension of the context-free grammar for strings. Since QUADSECTION is quite simple, there may exist the same or similar methods. However, the most important contribution of this paper is that it gives a guaranteed approximation ratio to the smallest grammar, which might stimulate further studies of improvements and extensions of grammar-based image compression.

Our proposed method has some similarity with *fractal image compression* [1,5]. Fractal image compression is based on a fact that parts of an image are often similar to other parts of the same image, and makes extensive use of these similarities. However, fractal image compression is usually computationally expensive. Furthermore, fractal image compression is usually lossy (i.e., it discards some information in the original image data). Different from fractal image compression, our proposed method is lossless and efficient, and has a guaranteed approximation ratio.

In this paper, we proposed a direct approach for grammar-based compression of image data. However, we can consider an indirect approach to compress image data in which a given image is first transformed into a string by means of raster scan and then is compressed using grammar-based compression algorithms for text data. Though it is difficult to extract patterns by such an approach, it might lead to better compression performances or better approximation ratios. Therefore, such an approach should be studied.

As shown in Section 6, for some types of binary or ternary image data, QUADSECTION had better performances than other standard image compression methods. However, in general, it is not better than those methods. In particular, QUADSECTION is not very useful for compression of gray-scale images or color images because QUADSECTION makes use of exactly repetitive patterns. Therefore, development of grammar-based compression methods for gray-scale images and color images is left as future work.

Acknowledgments. This work was partially supported by Grants-in-Aid #19650053 and #21700323 from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Bamsley, M.F., Demko, S.: Iterated function systems and the global construction of fractals. Proc. of Royal Society of London A399, 243–275 (1985)
2. Charikar, M., Lehman, E., Liu, D., Panigrahy, R., Prabhakaran, M., Sahai, A., Shelat, A.: The smallest grammar problem. IEEE Transactions on Information Theory 51, 2554–2576 (2005)
3. Drewes, F.: Grammatical picture generation: A tree-based approach. Springer, Heidelberg (2006)
4. Huffman, D.A.: A method for the construction of minimum-redundancy codes. In: Proceedings of the Institute of Radio Engineers, vol. 40, pp. 1098–1101 (1952)
5. Jacquin, A.E.: Image coding based on a fractal theory of iterated contractive image transformations. IEEE Transactions on Image Processing 1, 18–30 (1992)
6. Kieffer, J.C., Yang, E.H.: Grammar-based codes: A new class of universal lossless source codes. IEEE Transactions on Information Theory 46, 737–754 (2000)
7. Polvere, M., Nappi, M.: A feature vector technique for fast fractal image coding. Tech. rep., University of Salerno (1998)
8. Rytter, W.: Application of lempel-ziv factorization to the approximation of grammar-based compression. Theoretical Computer Science 302, 211–222 (2003)
9. Subramanian, K.G., Ali, R.M., Geethalakshmi, M., Nagar, A.K.: Pure 2d picture grammars and languages. Discrete Applied Mathematics 157, 3401–3411 (2009)
10. Ziv, J., Lempel, A.: Compression of individual sequences via variable-rate coding. IEEE Transactions on Information Theory 24, 530–536 (1978)

Effective Storage Structure for Multi-version XML Documents

Chang Chih-Chun, Cheng Ming-Shien, and Hsu Ping-Yu

National Central University, Department of Business Administration,
No.300, Jhongda Rd., Jhongli City, Taoyuan County 32001, Taiwan (R.O.C.)
984401019@cc.ncu.edu.tw,
mscheng@mail.mcut.edu.tw

Abstract. Office applications such as OpenOffice.org and Microsoft office are widely used to do everything you expect from your needs. Because of more and more requirements of information exchange and retrieve, XML becomes a standard in doing this way. With the adoption of using XML in both office application groups, the abilities for efficient storing historical office documents are become a growing issue. This paper introduces an efficient way to process multi-version XML documents. It is not only effective storage space need but also keeping the integral of original documents. It minimizes the change of data values or structures transmutation of historical XML documents. The purpose is to well-managed electronic documents for enterprises and all the messages were involved in should be preserved.

Keywords: OpenOffice.org, XML, Historical document, Storage structure.

1 Introduction

U.S National Archives and Records Administration (NARA) has mentioned in its strategic planning of 2006-2016, the current record of electronic document is in an exponential growth; therefore, it has to possess the competence of seeking, managing, using, sharing and properly handling[14] In 2006, two world-leading players of the office software, OpenOffice.org and Microsoft Office, have coincidently launched the data storage method with adopted the Extensible Markup Language (XML)[11] as the main body, and then XML has gradually become a standard format. XML is featured in with extensibility, structuralization and verifiability, which can not be restricted by computer platforms and programming languages; therefore, it has become one of those formats that recommended by World Wide Web Consortium (W3C)[15].

Impress is a presentation tool of the OpenOffice.org, Most presentation documents have showed high similarity among individual versions with only making slight adjustment in contents or the order of layout; therefore, the utilization of storage space is very inefficient, and the management of files has become complicated in the future.

The researcher has used the internal storage mechanism of OpenOffice.org to transform Impress into XML document, then made use of algorithm to integrate

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 249–266, 2010.

historical document with various versions into a single structure to avoid the repeated storage for document. Further, since the document format of OpenOffice.org is the Open XML document structure; therefore, we can also apply such format to all documents that conformed XML standards to achieve the goal of simplification.

In this XML study, the researcher has introduced the concept of threshold value for the first time. For the previous researches on processing the text or structure storage, the scope of data is restricted frequently. Thus, when there are more document formats that carried out the processing are based on XML, the processing standards will be definitely different in compiling and editing different electronic documents. Therefore, the application of threshold value can greatly increase the flexibility of document storage with responding to different requirements. This study use the processing of the threshold value and the application of algorithm to integrate XML historical documents with various versions into a single structure to achieve the goal of simplification. When recombining it into the document we need, the information of original version shall be presented without any omission.

2 Literature Review

This Section will explore and discuss relevant literature on the research objective of this study. Firstly, it will introduce Impress presentation software of OpenOffice.org, then it will explore the storage structure of XML document from many relevant literature, including the version control.

2.1 Brief Introduction to OpenOffice.org

Since the Impress is belonged to the office software, but its functions are not only used to process the daily business operations. It has comprehensive applications. Thus, it has characteristics that as compared with word processing software (Writer) and Spreadsheets software (Calc), and there existed great difference between them; therefore, if it can simplify the storage forms of files in the huge presentation document, and then it can make improvement in the preparation period of the presentation for internal meeting in the future. Or, it will have the great benefit to the information preservation level for external customers. Moreover, from the aforesaid causes, it may have easily caused the repeated storage for the document with the same topic.

2.2 Relevant Exploration of XML Document Storage

The design purpose of XML is to transmit and store data; thus, when most of XML document are processing, including the document modification, increase and delete and random access of nodes, and we called these actions were “Parsing”. There are two methods of Parsing: one is the Document Object Model (DOM) [16] and the other is Sequential Access XML Parser API (Simple API for XML, SAX) [12]. The framework of DOM is a standard that established by W3C, it has featured with the independence from languages and platform. When parsing the XML document, transformed the elements, attributes and texts of the document into a tree structure and store in the memory. Each node shall be regarded as an individual object and included the embedded value which can be operated programmers; in addition, a clear structure and easily understand will be its advantage. For SAX, it is s set of techniques that regarded XML document as the streaming interface, when downloading XML files with the sequential processing methods, and it will access the document with using the commands that set by programmers. However, it cannot be modified or accessed at will, as compared with DOM for using the tree structure to store in memory and occupied several folds of storage space for the original document, SAX is able to access any XML document regardless of its size, the designer's self-build model showed that when it only needs part of XML document, SAX will save even more storage space. Therefore, DOM and SAX are the Trade-Off XML Parser to each other, and they will make decision with regarding different applications[10].

[9] has also proposed the version control of XML, it will take the source XML data with based on the edit-based method to dismantle; however, this method needs to divided each element of this document into an individual object; however, the recover cost is too huge, thus S. Y. Chien et al. have proposed the usefulness-based clustering, which duplicated the storage in other pages. Even this process will be consumed some storage space, it still can make the recovering speed to become more efficient. However, this method is unable to solve the basic issue about controlling the traditional versions. And, to achieve the simplified storage space, it needs to dismantle the document of the source data into a minimum unit of each element, even it has increased the storage space with the clustering method to accelerate the recovering speed; however, the simplification of the primary space that still needs to be consumed more processing cost, thus, it showed an inefficient flexibility of processing document.

As for the latest literature about processing the historical version, they have emphasized on the combination with the data mining [4][5][6], within the ever-changing environment, XML data will be different by following different versions, data structures and texts will also be changed dynamically. Then, made use of the algorithm that similar to Apriori [7] and FPG[3] to mining the commonly changing part of the substructure of XML document [6]. From the dynamic XML version document to extract the slight change in time, and by means of these information to conduct the future forecast and application; and [5] thought that by following with the changes in versions, it will be existed a similar model among those substructures of XML document, as known as FCSPs, and it can sort out the longest FCSPs, from the resemble FPG algorithm; [4] has defined the changed substructure by following the changed version as the FRACTURE, then adopted the Level-wise and Divided-and-Conquer methods to mine the longest FRACTUREs; in addition, the mined result can be used to the future applications, such as the index or clustering, etc., of XML.

The recombination of this document is a area that received less researches [8], or only proposed those methods that shall be conducted the recombination but without any basis. Till the thought that made by[2], when querying the document, it shall show the necessary result of such user at last; therefore, it has to indicate any node and regard such node as the root node to conduct the document recombination on the stored data that adopted the relational database with using the sequential coding pattern, as well as increasing the querying efficiency.

3 Algorithm

3.1 Data Structure

The data structure is mainly composed of the coding techniques and cooperated with the threshold value to divide such document, and then preserve it in the form of matrix; among which, the access of original data is adopted SAX [12] to conduct the sequential access, and the coding has maintained the parent-son relationship for the tree structure, and conducted the processing in accordance with this basis [1].

3.2 Document Processing Algorithm

Algorithm is mainly divided into 3 parts: *f_list*, *XSS* and *Data_recovery algorithm.*

3.2.1 f_list Algorithm

As showed in Figure 1, the introduction of the threshold value concept, firstly, calculate the splitting value for document tags; among which, indicated the displayed texts, and English alphabet indicated the name of tags, if the name of tags are identical to each other, but their added attributes are different, then they shall still be regarded as the same tags. Then, such document has 12 tags, assumed A tags $=t_1$, B tags $=t_2,\ldots,L$ tags $=t_{12}$.Splitting value indicated that the level of tags for splitting such document. Therefore, as for the splitting value of the i_{th} tag, its formula will be $f_i = \frac{C(t_i)}{C(t_i)+S(t_i)}$, $C(t_i)$ is the total sum of i_{th} tag, and $S(t_i)$ is the total amount of sons and grandsons for t_i. Among which, the text will be regarded as a tag. For example, E tags' total amount is 3, and the total sum of its sons and grandsons tags will be 63; thus, the splitting value of E tags will be $f_5=3/(3+63)=1/66=0.0455$. If taking E tags to be the splitting point of this document, then such article can be divided into 4 parts, the sub-tree is composed of tags A, B, C and D, and 3 sub-trees that used E tags as the root elements. As for the splitting value of tags is greater than 0.35 的 tags, it has a power splitting capability, but it is easily to split articles into very scrappy; for example, the tag K's $f_{11}=13/29=0.4482$, tag L's $f_{12}=4/8=0.5$. If taking tag K or tags L as the splitting point, thus the document processing will cause a burden on such system, thus, we chose to enter the threshold

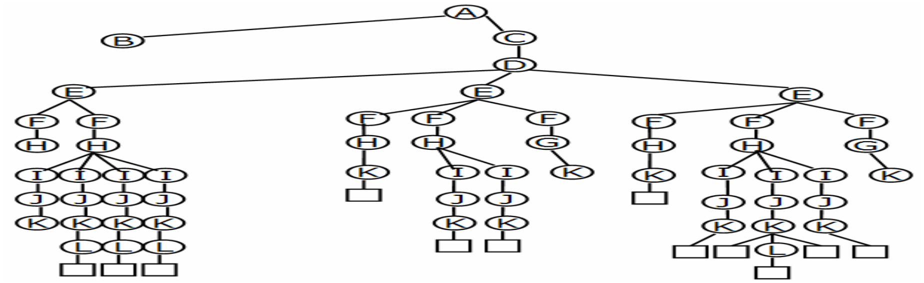


Fig. 1. Example of Impress Presentation XML Document Structure

value into the system, which needs to be entered into $\sigma_{\min}$ and $\sigma_{\max}$ respectively; among which, $0 \le \sigma_{\min} \le \sigma_{\max} \le 1$, for different types of document, it needs to decide the threshold value while the first version entering into the system. Among which, $\sigma_{\min}$ is the lower limit of the threshold value, that can remove those tags without splitting capability. $\sigma_{\max}$ is the upper limit of the threshold value; however, establish the suitable upper limit that can avoid articles from spilt too scrappy.

Therefore, if it assumed to set the upper and lower limits for the threshold value $\sigma_{\max}=0.25$ and $\sigma_{\min}=0.025$, and then the calculated splitting value is shown as follows:

Table 1. Tags Splitting Value

Tag	Splitting Value	Tag	Splitting Value
A (t_1)	$f_1 = \frac{1}{1+69} = \frac{1}{70} = 0.0143$	G (t_7)	$f_7 = 0.5$
B (t_2)	$f_2 = \frac{1}{1+0} = \frac{1}{1} = 1$	H (t_8)	$f_8 = 0.1176$
C (t_3)	$f_3 = \frac{1}{1+67} = \frac{1}{68} = 0.0147$	I (t_9)	$f_9 = 0.2195$
D (t_4)	$f_4 = \frac{1}{1+66} = \frac{1}{67} = 0.0149$	J (t_{10})	$f_{10} = 0.2813$
E (t_5)	$f_5 = \frac{3}{3+63} = \frac{3}{66} = 0.0455$	K (t_{12})	$f_{11} = 0.4482$
F (t_6)	$f_6 = \frac{8}{8+55} = \frac{8}{63} = 0.1270$	L (t_{13})	$f_{12} = 0.5$

Tag's splitting point is conformed to the threshold value, and tags are E, F, H and I, and the original XML document can be split into the sub-tree as follows:

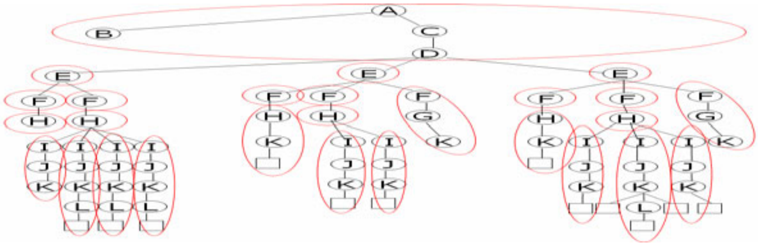


Fig. 2. XML Document Sectional Split

Thus, the detailed *f_list algorithm* is shown as follows:

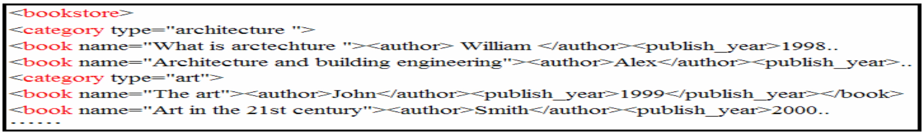
***f_list* algorithm** (Find Splitting Tags)

```
Input:  (1) $V_1$ (Version one of  XML document)
        (2) The threshold of splitting( $\sigma_{\min}, \sigma_{\max}$ ( $0 \le \sigma_{\min} \le \sigma_{\max} \le 1$ )
Output:  Splitting tags of XML document
Scan $V_1$ and $T=\{t_1, ,\ldots,t_n\}$ // T are the set of $V_1$'s all tags //
For each tag $t_i$ in T
  Count $C(t_i)$ and $S(t_i)$
    $f_i = \frac{C(t_i)}{C(t_i)+S(t_i)}$
```

```
        IF $\sigma_{\min} \leq f_i \leq \sigma_{\max}$
                THEN add $t_i$ to $F$ // $t_i$ is the splitting tag, and F is the set of all $t_i$ which
        splitting value match the threshold //
        End if
    Next
    Sort $F$ by splitting value ascending order into $f\_list$
```

3.2.2 XSS Algorithm

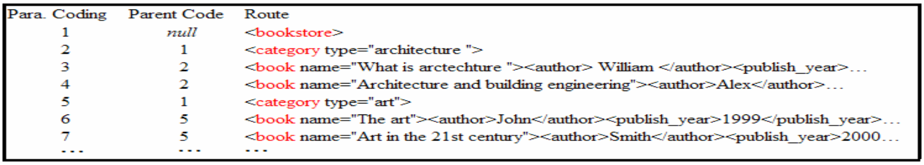
Makes use of the *f-list* algorithm to compute the splitting point for such type of XML document, and it is a pre-operating for the XML Storage Structure (XSS). *XSS algorithm* is adopted the splitting point for the *f_list of* XML document, and continuously carry out the splitting process on XML historical version document, XML document as shown in Figure 3, if assumed the splitting point is tags <category>and<book>, then such part of XML can be split as follows:



```
<bookstore>
<category type="architecture ">
<book name="What is arctechture "><author> William </author><publish_year>1998..
<book name="Architecture and building engineering"><author>Alex</author><publish_year>..
<category type="art">
<book name="The art"><author>John</author><publish_year>1999</publish_year></book>
<book name="Art in the 21st century"><author>Smith</author><publish_year>2000..
......
```

Fig. 3. Take <category> and <book> to be Splitting Points

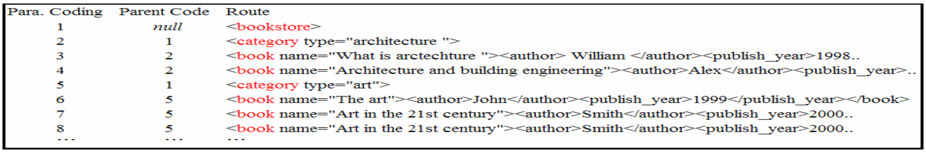
Since such document can split several word paragraphs; therefore, the selection of splitting point and threshold value will be the confirmation for the degree of easy modification of document; thus, when the upper limit of the threshold value is bigger, then the splitting degree of word paragraph will become more detailed. After processing such document, it will conduct the unique Paragraph coding for those word paragraphs that split out of each document, and then record the parent relationship with such word paragraph, and the relationship chart of this section, as shown in Figure 4:



Para. Coding	Parent Code	Route
1	*null*	<bookstore>
2	1	<category type="architecture ">
3	2	<book name="What is arctechture "><author> William </author><publish_year>...
4	2	<book name="Architecture and building engineering"><author>Alex</author>...
5	1	<category type="art">
6	5	<book name="The art"><author>John</author><publish_year>1999</publish_year>...
7	5	<book name="Art in the 21st century"><author>Smith</author><publish_year>2000...
...	...	...

Fig. 4. Joining Word Paragraph Coding

Tag <bookstore> is the root element of such XML document; therefore, its texts coding is 1, and since it is a root element, thus the field of ParentCode will be null, likewise, paragraph<category type="architecture” >’s texts coding is 2, and since such paragraph is the nested structure for the <bookstore> tags; thus, its parent paragraph will be coding 1<bookstore>, and so forth to complete coding. This relationship table of paragraph will be the import value of XSS algorithm. Before starting the algorithm processing, it has firstly conducted the transformation, to transform Parent Code into the matrix of subparagraph coding, and the reasons are as follows:

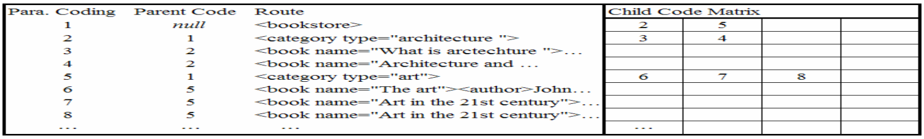


Para. Coding	Parent Code	Route
1	*null*	<bookstore>
2	1	<category type="architecture ">
3	2	<book name="What is arctechture "><author> William </author><publish_year>1998..
4	2	<book name="Architecture and building engineering"><author>Alex</author><publish_year>..
5	1	<category type="art">
6	5	<book name="The art"><author>John</author><publish_year>1999</publish_year></book>
7	5	<book name="Art in the 21st century"><author>Smith</author><publish_year>2000..
8	5	<book name="Art in the 21st century"><author>Smith</author><publish_year>2000..
...	...	...

Fig. 5. Another Possible Cutting Situation of XML Document

Assumed the relationship table that dissolved from the original XML document as shown in the aforesaid figure, then, in fact, coding 7 and 8 have the same text contents, in order to achieve the high-efficient simplified XML document, the coding 8's word paragraph will be removed; however, when conducting the document recover, the information of such word paragraph will be no longer appeared. Thus, after the paragraph relationship table imported *XSS* algorithm, it needs to transform the subparagraph coding matrix firstly. Assumed r_y indicated every word paragraph, thus rc_y and rfc_y are the corresponding coding and parent paragraph coding for r_y; for example, r_2 is <category type="architecture" >, and rc_2=2 and $rfc_2=1$ are the corresponding codes. After transformed the subparagraph coding table into a matrix, as shown in Table 2. In such matrix, $rcc_{x,y}$ indicated the subparagraph coding, x, y are indicated the horizontal coordinates and vertical coordinates respectively, values of y shall be corresponding to the word paragraph of the paragraph relationship table; for example, r_5 is <category type="art">, thus $rcc_{1,5}$ =6, and its first word subparagraph is r_6, likewise, $rcc_{2,5}$ =7, $rcc_{3,5}$ =8. If such word paragraph is leaves node, which meant, except parent paragraph, that it never have the subparagraph, as shown in the figure, the subparagraph coding $rcc_{1,y}$, r_3, r_4 , r_6 , r_7, and r_8 is null (empty value).

Table 2. Transformation of Subparagraph Texts Coding



Para. Coding	Parent Code	Route	Child Code Matrix			
1	*null*	<bookstore>	2	5		
2	1	<category type="architecture ">	3	4		
3	2	<book name="What is arctechture ">...				
4	2	<book name="Architecture and ...				
5	1	<category type="art">	6	7	8	
6	5	<book name="The art"><author>John...				
7	5	<book name="Art in the 21st century">...				
8	5	<book name="Art in the 21st century">...				
...	...	...	...			

```
Sub Convert ( $v_i$ )
  For each $r_v$ in $v_i$ paragraph relation table where it hasn't converted
    Using $rc_v$ and $rc_v$ to generate Child Code Matrix
    If any $r_v$ doesn't have child then
      $rcc_{1,v}$ is empty
    End if
  Next

End Sub
```

After transformed the subparagraph coding into a matrix, r_7 =r_8=<book name="Art in the 21st century">.., by means of the *XSS algorithm* processing, r_8 word paragraph and its corresponding coding will be cleared out, but the document's information has not lost yet, and it only needs to change $rcc_{3,5}$=8 into $rcc_{3,5}$=7. Thus, during the period of document recovery, it still can express the information of three word paragraphs of r_5.

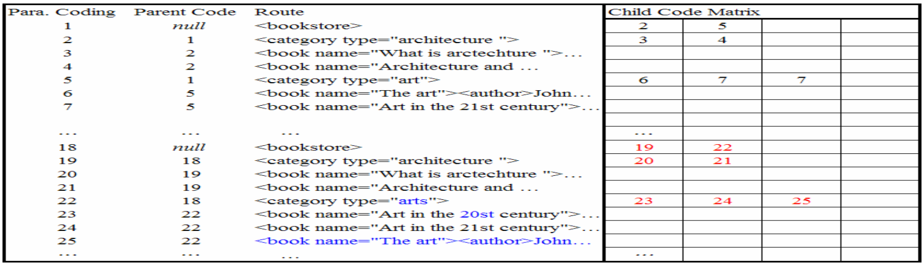
Before carrying out the *XSS* algorithm, firstly define *R* is an assembly set that formed by r_1~ r_{y-1}. Set *R* is the text comparing basis of comparing with r_y. Where $RCC_y=\{ rcc_{1,y} ,rcc_{2,y} ,.....rcc_{x,y}\}$, and $RCC_y \subset RCC$; therefore, *XSS* algorithm is shown as follows:

***XSS* algorithm** (XML Storage Structure)

```
Input:  $v_1$, $v_2$, $v_3$ .... $v_n$ (paragraph relation table of history version of XML document
        from f_list splitting
   Output:          XML Storage Structure (XSS)
   Call Sub Convert ($v_i$ )
        For each $rcc_{1,v}$ which is empty
                Compare $r_v$ with R set
                If $r_v$ matches any $r_{v-n}$ |n:1<=n<y then
                Find $rcc_{x.k}$ = $rc_v$
                Modify $rcc_{x.k}$ to $rcc_{v-n}$
                End if
        Next
        For each $rcc_{1.v}$ which is not empty
                For each $col_x$ in Child Code Matrix
                        If $r_v$ = $r_{v-n}$ | n:1<=n<y and
          { $rcc_{m.v}$ , $rcc_{m.v-n}$ } |m=1~x, $rcc_{m.v}$ , $rcc_{m.v-n}$ then
                                Find $rcc_{x.k}$ = $rc_v$
                           Modify $rcc_{x.k}$ to $rcc_{v-n}$
                        End if
                Next
        Next
   End if

Clear every $r_v$, , $rc_v$, , and $rfc_v$ and $rcc_{m.v}$ | m=1~x which $rc_v$ is not root number and $rc_v$
doesn't exist in RCC
```

For example, as shown in Table 3, since the parent paragraph coding of these two word paragraphs have *null* value, thus it indicated that there are two documents; among which, at the time when XML document of v_1 has initially entered into the system, the word paragraph r_7 of coding 7 will be deleted by adopting the algorithm processing. After document of v_2 entered the system, firstly, it shall split the XML document of v_2 by using splitting point *f_list*, and store the split document, together with its texts coding paragraph relationship table into the system, as shown in the left bottom of Table 3. Among which, difference existed in 3 points for these 2 versions when comparing Version 1 and Version 2. Changes similar to r_{22} that can be called as the structural variation. Changes of r_{23} are the revisions of texts, and r_{25} is indicated the movement of word paragraph, which moved from the first nested structure of the original <category type="art"> to the third, but there is no such correction or modification made for the text contents.

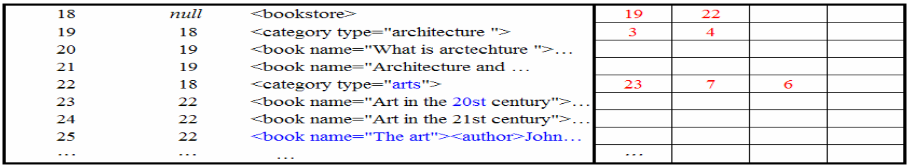
Table 3. When Version 2 Document Entering the System



Para. Coding	Parent Code	Route	Child Code Matrix			
1	*null*	<bookstore>	2	5		
2	1	<category type="architecture ">	3	4		
3	2	<book name="What is arctechture ">...				
4	2	<book name="Architecture and ...				
5	1	<category type="art">	6	7	7	
6	5	<book name="The art"><author>John...				
7	5	<book name="Art in the 21st century">...				
...	...	...	...			
18	*null*	<bookstore>	19	22		
19	18	<category type="architecture ">	20	21		
20	19	<book name="What is arctechture ">...				
21	19	<book name="Architecture and ...				
22	18	<category type="arts">	23	24	25	
23	22	<book name="Art in the 20st century">...				
24	22	<book name="Art in the 21st century">...				
25	22	<book name="The art"><author>John...				
...	...	...	...			

By means of the data explanation, it begins to enter the algorithm of XSS. After document entered the system, it will immediately call up the subprogram, Transform (v_2), the newly added subparagraph coding matrix as shown in Table 3. When starting the simplification process to the word paragraph of v_2, it shall be split into 2 parts, the first part is that when the subparagraph coding $rcc_{1,y}$ of the word paragraph r_y is null value, as shown in the above figure, r_{20}, r_{21}, r_{23}, r_{24} and r_{25}, conducted the comparison with these word paragraphs and the set of $R_y = \{ r_1 , r_2 \ldots r_{y-1} \}$, r_{20} will compared with r_1 , $r_2 \ldots r_{19}$ for texts, if those compared results are different to each other, then it can regard such word paragraph as the newly added part of the historical version, or the part that has been modified or revised, such as r_{23}; if the compared results are matched, such as r_{20} is identical to r_3, then search the texts coding “20” of r_{20} under the subparagraph texts coding, and changed such subparagraph texts coding into the texts coding “3” of r_3, namely $rcc_{1,19}$=3; at the same time, as for the moving part of texts, such as r_{25}, since its texts have not yet changed or modified, thus, the compared results are identical to each other; therefore, according to the same method, $rcc_{3,22}$=3 will be changed into “6”, and v_2 is changed as shown in Table 4.

Table 4. Step 1: When $rcc_{1,y}$ is a Null Value



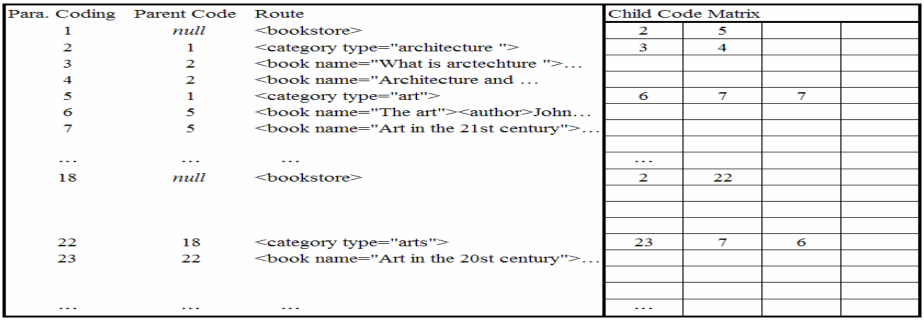
18	*null*	<bookstore>	19	22		
19	18	<category type="architecture ">	3	4		
20	19	<book name="What is arctechture ">...				
21	19	<book name="Architecture and ...				
22	18	<category type="arts">	23	7	6	
23	22	<book name="Art in the 20st century">...				
24	22	<book name="Art in the 21st century">...				
25	22	<book name="The art"><author>John...				
...	...	...	...			

After comparison, when $rcc_{1,y}$ equals to null value, it can enter to the second part, $rcc_{1,y}$ not equals to the null value, as r_{18}, r_{19} and r_{22}in the above figure, the texts comparing methods are also as abovementioned; however, except the same methods for texts, the subparagraph texts coding shall be totally identical, such as r_{19} and r_2, their texts are identical to each other, as well as $rcc_{1,19} = rcc_{1,2}$ and $rcc_{2,19} = rcc_{2,2}$;

therefore, it can be showed that there is no change in structure or text, and it can look for the subparagraph coding “19” and change into the coding “ 2” of r_2.

Through Step 1 and Step 2, it can immediately found the repeated part out of v_1 and v_2 in XML document, if it needs to achieve the structural simplification, it only needs to make each word paragraph coding rc_y not become the root word paragraph, or when the subparagraph texts coding is not appeared, delete the data for entire row, including r_y, rc_v , $rfc_{y,}$ and $rcc_{m,y}$|m=1~x, , and only retain the newly added part and changed/ modified part. The completely simplified XXS is shown in Table5.

Table 5. Simplified XSS Structure



Para. Coding	Parent Code	Route	Child Code Matrix			
1	*null*	<bookstore>	2	5		
2	1	<category type="architecture ">	3	4		
3	2	<book name="What is arctechture ">...				
4	2	<book name="Architecture and ...				
5	1	<category type="art">	6	7	7	
6	5	<book name="The art"><author>John...				
7	5	<book name="Art in the 21st century">...				
...	...	...	...			
18	*null*	<bookstore>	2	22		
22	18	<category type="arts">	23	7	6	
23	22	<book name="Art in the 20st century">...				
...	...	...	...			

3.2.3 Data_recovery Algorithm

The main purpose of the texts structure for XSS is to retain and preserve the information of the original XML historical document, except for that the repeatedly appeared part can be deleted as well with simplifying the storage space for such document. When we need the information of this document, it can be then presented immediately; therefore, it needs to develop algorithm to recombine the words and paragraphs. Continuously make use of the XSS structure in Table 5 to describe the process of document recovery. If users need the Version 2 XML document, he/she shall enter the version number $j=2$ into the system, through the algorithm to determine the coding of Version 2 root word paragraph; in this example, it will be rc_{18}=18 and using the recursive method to process rc_{18}. Since the texts coding of the subparagraph is represented by matrix, thus, it will be adopted the Depth-First Search (DSF) method to recover the document.

When entering rc_{18} into the recursive algorithm, firstly to process the first subparagraph texts coding , the conditions of determination have the following 2 steps sequentially: first step, to recombine word paragraph into the document, such as when rc_{18}=18, duplicate and store <bookstore> into the XML file; in addition, transmitting the first word subparagraph coding $rc_{1,18}$=2 into its own recursive algorithm, and concurrently store the r_2 word paragraph<category type="architecture">, and so forth. The second step is to determine whether $rcc_{x,y}$ is a null value or not; if so, just exit the recursive algorithm then. Among which, the null value has indicated two meanings, one is indicated the searching that has completed to the leaves word paragraph, as $rcc_{1,23}$ in 0. Another situation has completely searched all subparagraph texts coding, such as $rcc_{3,18}$, $rcc_{4,22}$,. *Data_recovery algorithm* is shown as follows.

***Data_recovery* Algorithm** (XML document recompose)

```
Input:      XML Storage Structure (XSS)
Selected historical XML document version number $j$
Output:     Original XML document $v_i$
Main ( )
Find $v_i$ root number $rc_v$
Recursive ( $rc_v$ )
Return $v_i$
Function Recursive ( $rc_v$ )
For each $col_x$
    $paragraph\_n = cc_{x..v}$
    Copy and store $r_{.v}$ into XML document when entering function
    If $rcc_{x,y} \in \phi$ then
        Exit for
    End if
    Recursive ( $paragraph\_n$ )
Next
```

4 Empirical Analysis

This section is mainly composed of 2 major parts: first part is to describe the data-oriented slides image file of the OpenOffice.org Impress presentation; and the second part is to analyze the results and data that came from those experiments.

4.1 Experiment Design

The data of this experiment is the slides that used to make presentation of OpenOffice 2.0 in **Ce**ntrum der **B**üro-und **I**nformations**t**echnik (CeBIT) on March 2006, the document contents included graphs, difference in font sizes, color usage, hyperlink and animation layout, and the total page number is reached to 37 pages, which is identical to the slide quantity of common presentation. Since the languages is German, thus, in order to make it to be easily understand, this experiment has transformed the original Germany version Impress document into the English version [13]; in addition to make such translation, the rest parts of this presentation have not yet changed and maintain the framework of the original document.

Besides, in order to increase the document's variability, selected 102 key words from those articles and randomly changed key words to assist in yielding different versions document.

4.2 Experiment Result and Analysis

Since the source data is only one part existed, thus it shall use the simulative methods to create different versions. Among those possible causes of yielding different versions, the presenter's preference will be different for each other, thus it cannot be simulated; and another factor is that the continuously increased page number while preparing the presentation slide file; as a result, based on such assumption to design the Experiment 1.

Experiment 1: Continuously Increasing Page Number for Each Version Document Experiment Description: when simulating the users with compiling document, the process of gradually increasing the contents for such presentation. Since it only has a copy of the original text with a total page number is 37; therefore, it assumed that the OpenOffice Impress of v_1 only has 8 pages, and this part that may not changed ever will be completed in advance; therefore, v_1 is contained the first 4 pages and ending 4 pages of the original document. Each version will be increased one page more than the previous one; thus, v_{30} is the original content, and the experimental result is shown as in Figure 6:

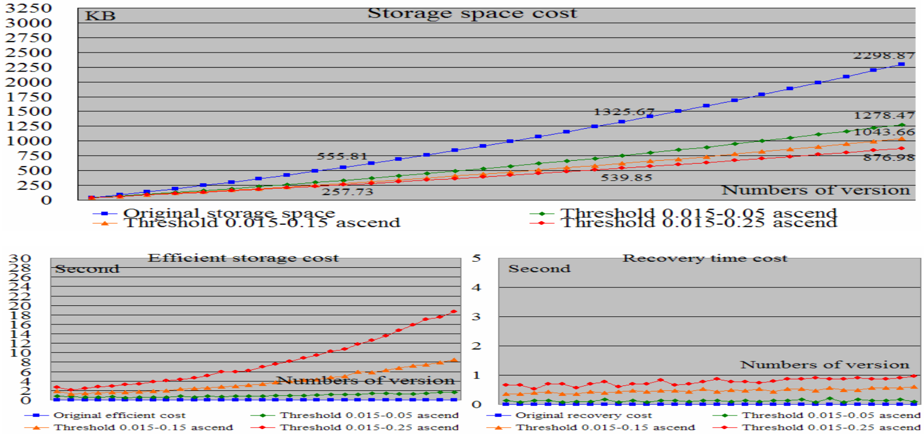


Fig. 6. Document Pages Number Increasing Storage Space, Simplify and Recovery Time

Analysis on experimental result: the original storage space with accumulated 30 versions is 2298.87 KB, and if use the example of 0.015-0.25 to make explanation, in v_{10}, it may possibly save about 54% of storage space, and v_{20} will be about 59%; after completed 30 versions, it can be simplified about 62%; thus, the document will be saved more space (storage) by following the evolution of versions. Among which, the upper curved graphs showed that when the storing the original document contents, it shall have the non-preset effectiveness; such as the changes in font sizes and colors, and the increase in animation, etc., they will be made definition in XML, when increasing 1 page, except the storage of page, it still needs to contain the definition space for the special effect; in addition, when using the *XSS algorithm* to process, it will delete the repeated effect definition between versions, and the degree of upper curved will be slowed down, and the more versions will make more optimal rate of simplified space.

In addition to compare with the original file and the storage space after used the algorithm to make process, at the same time, the threshold value may also be changed as well, and then inspect whether there is any difference in the usability of space or not, as shown in Figure 6, the threshold values of the rhombus points are among

0.015-0.05; comparatively, the presentation document is adopted “pages” as the standard of minimum division. 0.015-0.15 are based on the “text boxes”, sphere points are around 0.015-0.25, which is indicated the text of each line, or namely the item sign, Bullet, as the minimum division standard. From the figure, there is only 0.1 for the difference in the upper limit of the threshold value, but the result will be greatly different. When changed the value of upper limit from 0.05 into 0.15, the improved efficiency will be more and the efficiency of its original simplification is not reached 50%. In this experiment, it has assumed that the version document is belonged to the increasing pattern, there is no change in document’s contents, but the page number is increasing by the method of inserting page.

As for the time comparison, since the most original un-processed documents that don’t need time to conduct the simplified and recovered storage space; therefore, it assumed that consumed zero second. Within the different threshold values for other lines, and the time that system consumed for processing, and this experiment increasing version document,; thus, when the upper limit of threshold value is bigger, then the number of word paragraphs of such version will be dividend into a bigger number then. In addition, when carrying out the simplification for the structure, it will be consumed more computing capability of computer.

Experiment 2: Fixed Page Number among Versions and Replacing Key Words with Texts

Experiment Description: Since the reporting targets are different, or with same topic but different words and phrases that different presenters favorably used in their individual presentation. This experiment has acquired 102 key word from 37 pages of original presentation file, and each key word cab be paired with 2 synonyms; for example, languages can be replaced by vernaculars, or changed to tongues; on the contrary likewise, this experiment also has 37 pages but different key words from historical document, the graph and analysis are shown in Figure 7.

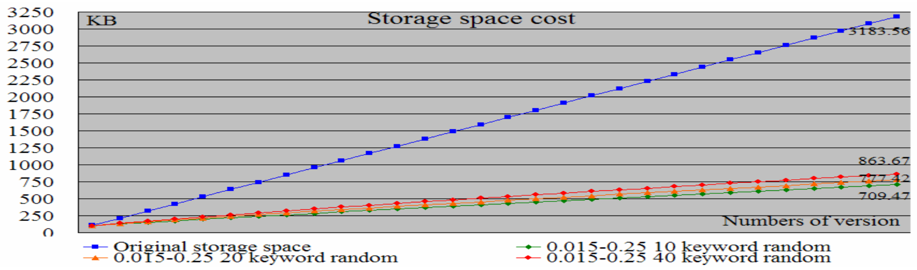


Fig. 7. Pages Number Fixed and Replace Key Words with the Text of Storage Space

Extensible Experiment 1, since the ratio of space saving for these two types of threshold values 0.015-0.15 and 0.015-0.25 will be more than 50% for both types; therefore, in Experiment 2, it will be listed the splitting value firstly, as 0.015-0.25 to be the basis of saving space. The following figure is showed the comparison with another threshold value 0.015-0.15.

Experiment Result Analysis: from the experimental result, even when the threshold values is 0.015-0.25, and document contents have randomly replaced with 40 key words, *XSS* algorithm has saved the space about 73% with a significant saving effectiveness on document. From Figure 7 , we founded a phenomenon. When the frequency of replacing key words has changed from 10 into 20, the space that consumed additionally will be 67.95KB, and the number of replacing key words will be increased another 20, then it will be consumed about 86.25KB for the storage space. It just has a slight difference between 10 and 20 replacing numbers, since the threshold value 0.015-0.25 has divided on the basis of Bullet; therefore, it will not be sensitive to the changes in key words in such document; however, if it has adopted the text boxes of the threshold value 0.015-0.15 to divide, the comparative result is showed in Figure 8:

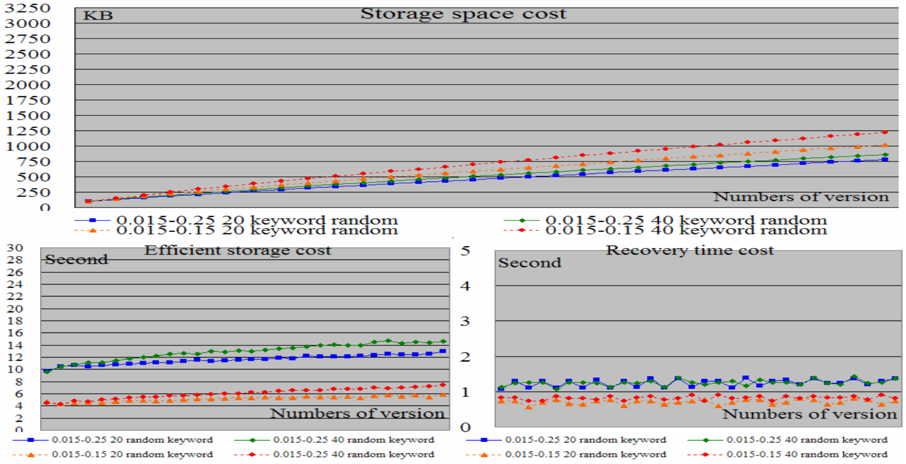


Fig. 8. Comparison with Storage Space, Simplify and Recovery Time between Different Threshold Values

Figure 8 has showed a significant difference that caused by comparing with these two threshold values. When selecting the threshold value with bigger upper limit value, it will be obtained a more simplified space since the threshold value 0.015-0.15 is adopted text boxes as the basis of splitting. Thus, if a text box contained 4-line texts, which meant 4 item signs, when changed one line of its texts, then it will be regarded as a different version from different text boxes contents of different versions; therefore, the system will prepare one more portion of text boxes than storage. In fact, other 3 lines texts are completely identical to each other, which has resulted a lower efficiency on simplification, and it is clearly showed in this experiment with more than 40 times of replacement. However, if it has adopted 0.015-0.25 as the splitting standard for each line of texts, and changed the key words will only resulted in replacement of the texts for such line but it will not completely influence the texts in the same box; in addition, it can adopt algorithm to delete the repeated part of such texts, thus, it will not store in the unnecessary area.

The time of simplification in this experiment will be slightly increased on the linear basis; and the time difference between these two threshold values will be about 2 folds; in addition, in the experiment of recovering the document, the threshold value with fine division will take one second more than the other and there is no change in pages number; thus, the time will be kept on horizontally steady without showing the increasing or lowering trend.

Experiment 3: Pages Number is Increasing and Replacing Texts with Key Words
Experiment Description: in the aforesaid experiment, it has simulated the single situation only; therefore, the Experiment 4 will simulate the situation that will be occurred more frequently everyday – by following the versions, users may have also increased the pages number, as well as modified the original text contents at the same time. Therefore, as for the data simulation, it will adopt the method of inserting 1 page into each version, that is, 8 pages for v_1, and 37 pages for v_{30}, as to the modification of the previous/ original text contents, it has also adopted the method of replacing with 10, 20 and 40 key words, respectively, to complete 30 historical versions document in order to inspect the coexistence for these two situations. Since in Experiment 1 and Experiment 3, regardless of pages number increasing or replacing key words, the threshold value 0.015-0.25 has an excellent ratio of saving the space; thus, this experiment is divided on the basis of the threshold value, as shown in Figure 9.

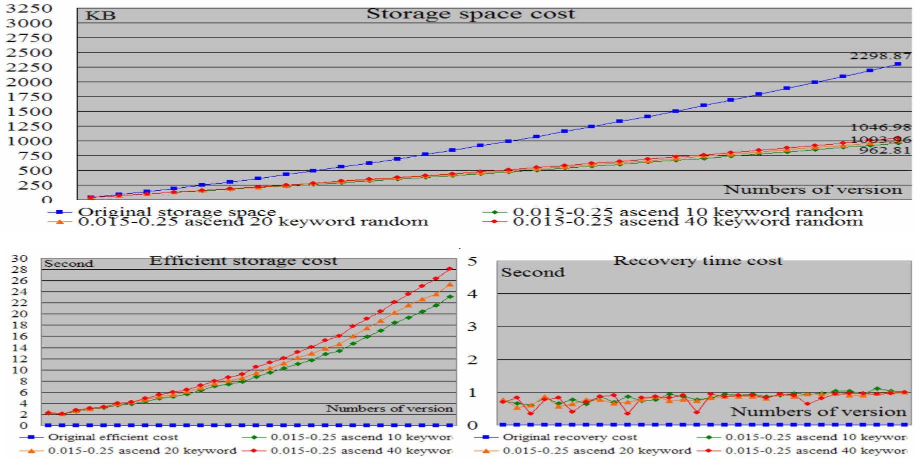
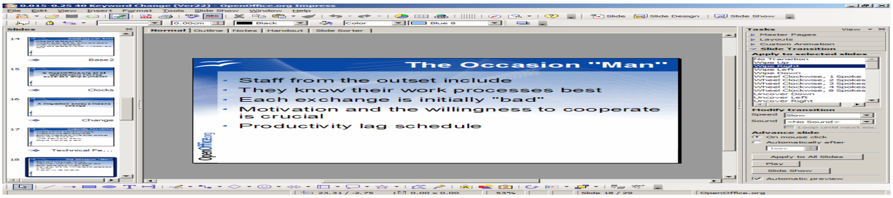


Fig. 9. Pages Number is Increasing and Replacing Key Words for Storage Space, Simplify and Recovery Time

Experiment Result Analysis: as shown in aforesaid figure, it can find out an interesting phenomenon. When increasing the number of replacing key words from 10 to 20, and the storage space of file has increased a margin as same as increasing the number from 20 to 40; thus, it may changed back to the original words or phrases at the end of the process; thus, the systems needs to maintain the first time it appeared and to simplify the repeated part. Therefore, the trend of such phenomenon can be roughly seen, when replacing more key words, or there are more versions, and increasing repeated part; then the saved space is more efficient. As compared with the time that will be identical to the aforesaid experiment, the time of simplification will be depended on the number of data for such version, and its recovery time will maintain stable.

Experiment 4: Document Response

Experiment Description: this experiment is attempted to present that even after completed the simplification of document, the remained information is still able to make response at all time. This experiment is constructed at the threshold value 0.015-0.25 and replacing with 40 key words, which used to verify whether it would be recovered or not. The eighteenth page of v_{22} and the first page of v_{10} can be randomly selected and recovered as the screen as follows:

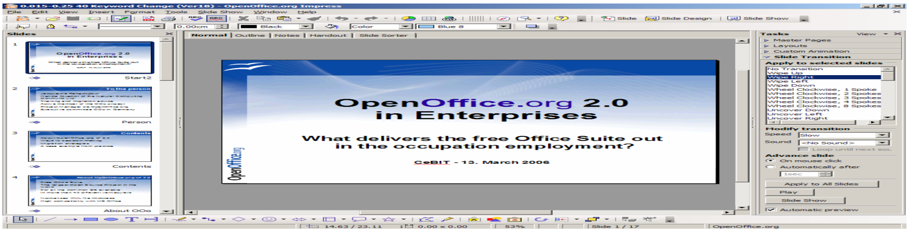

0.015-0.25 40 Keyword Change (Ver22) - OpenOffice.org Impress

The Occasion "Man"

- Staff from the outset include
- They know their work processes best
- Each exchange is initially "bad"
- Motivation and the willingness to cooperate is crucial
- Productivity lag schedule

Fig. 10. Recover the eighteenth Page of v_{22}

Experiment result (1): through the eighteenth page of v_{22} as shown in Figure 10, after recovered, it included the ground colors, titles and the text inside the text boxes that will remained the size of the original version, as well as the setting of animation. The title of the original source presentation document [13] is The Factor "Man", after randomly replaced 40 key words through the entire document, and the contents of this page have not been changed, but the title has replaced to The Occasion "Man"; therefore, from the screen of this experiment result (1), when recovering document, it has also maintained the information of the original document, and after compared the replaced document with the original document, there is no difference between them at all.


0.015-0.25 40 Keyword Change (Ver18) - OpenOffice.org Impress

OpenOffice.org 2.0
in Enterprises

What delivers the free Office Suite out in the occupation employment?

CeBIT - 13. March 2006

Fig. 11. First Page of Recovered v_{10}

Experiment result (2): as compared the version document of v_{10} with the original source document [13], the title of OpenOffice.org 2.0 in Enterprises has not been replaced yet; however, the contents, carries and delivers of another text box have been replaced, and replaced business with occupation. In addition, in this experiment, the result, color, font, animation of the recovery document can be normally presented. As compared with the original v_{10}, the result will be identical to each other, loyally presented the version of replaced key words; thus, it has verified that those experiment are successful.

5 Conclusion and Future Research Suggestions

5.1 Conclusion

In this study, it has practically adopted the OpenOffice.org Impress presentation document to carry out the experiment, and utilize algorithms to simplify the contents of XML files in Impress. From those experiments to simulate the time when presenters making the slides, the situation of continuously increasing and deleting the document; however, under the condition of continuously changing in the structure of XML document, if it can reach to 30 versions, then it can save about 62% storage space. In addition, it is assumed that users have continuously changing the key words or sentences of the document, it can be clearly expressed even the scope of modifying key words is increased; however, the storage space may still be able to reach an excellent level. At last, it has conducted the Experiment 4 which has conformed to the actual situation, even its changing scope of the document is the maximum, and algorithms still can acquire identical part from files to make simplification. From those experiments we can understand that the establishment of threshold value, and make users to flexibly use the storage space, in accordance with each document with various topics, to establish different threshold values to optimize the storage space.

In the last experiment of this study, by means of integrating 30 versions of Impress presentation document into a single system, it can not only express to practically recover the historical document, but also conduct the optimal management to documents:

5.2 Future Research Suggestion

This study is mainly focused on the processing of XML open structure document storage, especially it will be developed its advantage over the office document with various versions; in addition, by means of the experimental simulation of the possible occurrence to verify this algorithm is feasible.

The processing of storage space in the future that shall not be restricted to texts only, graphs may frequently appeared around us, and the structure of graphs may possibly evolved in the future; for example, the chemical molecular bonding, etc. When the graph needs to be stored, these version-evolved graphs in the space of computer memory is complicated and useless, but they will be preserved as the historic message; therefore, if this method can be promoted to graphs, and recorded the parent-son and evolving relationships for such combination that will make graph to be well-managed as well.

Consideration factors of version, except texts and graphs, in the practical conditions, as for those industries that will extremely focus on their storage space, such as the retail business or manufacturing industry that needs bulk sales and purchase the components and parts, in order to be well-stored the transaction data, those factors cannot store the most detailed data into the database, but needs to regularly adopt the tape to make such storage. It will take a large quantity of time for the sequential access process when responding historic data; however, most of the historic data are similar and only the time points of occurrence are different. Thus, in the future, it only needs to use database to store the transaction data, and further it can be developed as a tool of data analysis in order to help with the business competitiveness as well.

References

1. 林昌正,「多XML文件整合萃取工具之研究」, 國立中央大學, 碩士論文, 民國97年。
2. Chebotko, A., Liu, D., Atay, M., Lu, S., Fotouhi, F.: Reconstructing XML subtrees from relational storage of XML documents. In: Proceedings of the Second IEEE International Workshop on XML Schema and Data Management (XSDM 2005), in conjunction with ICDE 2005, Tokyo, Japan (April 2005)
3. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD 2000), Dallas, TX, pp. 1–12 (May 2000)
4. Chen, L., Bhowmick, S.S., Chia, L.T.: FRACTURE-Mining: Mining Frequently and Concurrently Mutating Structures from Historical XML Documents. Elsevier Science Journal: Data & Knowledge Engineering 59(2), 320–347 (2006)
5. Chen, L., Bhowmick, S.S., Chia, L.T.: Mining Maximal Frequently Changing Subtree Patterns from XML Documents. In: Proceedings of the 6th International Conference on Data Warehousing and Knowledge Discovery (DaWaK), Zaragoza, Spain, pp. 68–76 (2004)
6. Rusu, L.H., Rahayu, W., Taniar, D.: Mining Changes from Versions of Dynamic XML Documents. In: Nayak, R., Zaki, M.J. (eds.) KDXD 2006. LNCS, vol. 3915, pp. 3–12. Springer, Heidelberg (2006)
7. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proc. 1994 Int. Conf. Very Large Data Bases, VLDB 1994, Santiago, Chile, pp. 487–499 (September 1994)
8. Krishnamurthy, R., Chakaravarthy, V.T., Kaushik, R., Naughton, J.F.: Recursive XML schemas, recursive XML queries, and relational storage: XML-to-SQL query translation. In: Proc. of the ICDE Conference, pp. 42–53 (2004)
9. Chien, S.Y., Tsotras, V.J., Zaniolo, C.: Efficient schemes for managing multiversion XML documents. VLDB J. 11(4), 332–353 (2002)
10. Wang, F.J., Li, J., Homayounfar, H.: A space efficient XML DOM parser. Data & Knowledge Engineering 60, 185–207 (2007)
11. Extensible Markup Language (XML), `http://www.w3.org/xml/`
12. Megginson Technologies: Simple API for XML, `http://www.megginson.com/downloads/SAX/`
13. OpenOffice.org 2.0 in Enterprises. English version, `http://www.ba.ncu.edu.tw/dmerplab/CeBIT_OOo_En.odp`
14. Preserving The Past To Protect The Future, The Strategic Plan of The National Archives and Records Administration (2006-2016)
15. World Wide Web Consortium, `http://www.w3.org/`
16. W3C's Document Object Model (DOM), `http://www.w3.org/DOM`

An Ontology-Based Platform for Scientific Writing and Publishing

Hao Xu[1,2,3], Changhai Zhang[1], Yang Ai[1,3], Ziwen Wang[1], and Zhanshan Li [1,*]

[1] College of Computer Science and Technology, Jilin University,
Qianjin Street 2699, Changchun 130012, China
[2] Center for Computer Fundamental Education, Jilin University,
Qianjin Street 2699, Changchun 130012, China
[3] Department of Information Engineering and Computer Science, University of Trento,
Via Sommarive 14, Povo 38123, Italy
hao@disi.unitn.it, changhai@jlu.edu.cn,
aiyangjilin@yahoo.com.cn, mierkelin@sina.com,
zsli@email.jlu.edu.cn

Abstract. Writing scientific discourses and publishing academic results are integral parts of a researcher's daily professional life. Although tremendous magic have been brought by advancement of digital library technologies and social networking services, there are still no off-the-shelf utilities for strategic writing, reading and even publishing. In this paper, we propose an ontology-based platform for scientific writing and publishing featured with its semi-automatic metadata generation and semantic-linked composition, aiming to facilitate the efficient creation, dissemination and reuse of scientific knowledge.

1 Introduction

Since the Elsevier's Journal *Cell* launched the new online publishing format "Article of the Future" [1] in the year of 2010, a revolution in scientific publishing has taken place again. In recent years, a handful of models were proposed for scientific discourse representation grounded on the Rhetorical Structure Theory (RST) [1] such as ABCDE Format [2], SALT (Semantically Annotated LaTex) [3], etc. Also, several applications of ontology-aware text mining and retrieval systems, e.g. Textpresso [2] and iHOP [3] are widely accepted in particular domains of biological science.

Ontology-based structure for representing scientific papers, along with semantic web techniques for linking related entities and concepts, forms the basis for new types of scientific writing and publishing. Within the construction of our proposed ontology, we also consider defining the patterns of papers' logical structuring for strategic writing, reading, and the interoperability with other domain ontologies as well as metadata schemes. The advantage of such proposition also evolves the

[*] Corresponding author.

[1] "Article of the Future": http://beta.cell.com/

[2] Textpresso: http://www.textpresso.org/

[3] iHOP: http://www.ihop-net.org

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 267–271, 2010.

collaborative operations in the research communities including metadata annotation, maintenance, sharing and so forth.

In this paper, we briefly introduce our ongoing project [4, 5, 6, 7] of an ontology-based platform for scientific writing and publishing. The objective of this project aims at applying rhetorical structure theory to online authoring, navigating and annotating. Thus readers can access to their own interests directly without being overwhelmed by the extra noisy information. The previous indivisible linear-structured papers are re-organized by ontology-based structure within meaningful semantics.

2 Overview

This platform consists of two main components. One is an *Editor* used for semantically authoring, where "semantically" means that the system may offer the functionality of encoding various semantic links. For instance, in traditional system one is difficult to track versioning during the manuscript's evolving. Semantic links do definitely solve these problems via sets of URLs and metadata information. Moreover, various reviews and comments associated to different draft versions are also considered when the Editor is designed. We call it lifecycle management which is one of functions in the Editor. Others services like local metadata, writing patterns, strategic reading, ubiquitous reuse and semantic search are supported.

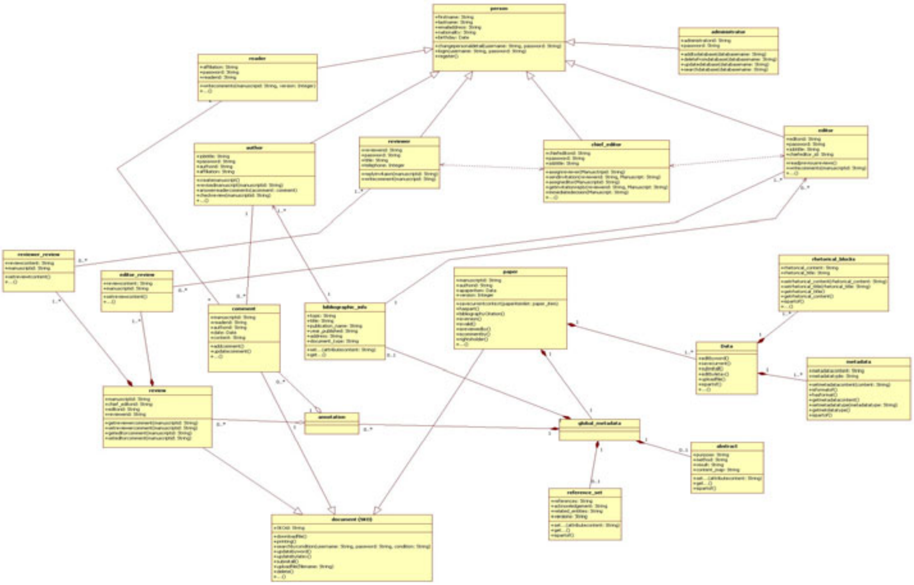


Fig. 1. UML Class Diagram for Ontology Development

The other one is an *Online Portal* for navigating. More rhetorical structure templates go beyond the domain-specific prototypes of existing online publishing systems. Types of document and roles of person are assigned specifying metadata schemes for faceted search. Social networking makes the community collaborative and communication efficient.

Herein, an overview of platform design is illustrated in Figure 1. This UML class diagram shows the core of ontology development, which is as well as the core of whole platform. Document (Paper, Review, and Comment) and Person (Author, Reader, Editor, Chief Editor, and Administrator) constitute the main classes, along with sets of attributes and operations. Sub-patterns are implemented for each types of document, i.e. paper, review and comment, likewise. This methodology aims to divide linear papers into parts, namely rhetorical chunks. Instead of ordered sections and subsections, there rhetorical chunks are connected via types of semantic links mentioned above. One paper is composited by Global Metadata and Data, while Data is organized by several rhetorical blocks and multiple local metadata related with each rhetorical block. All metadata can be accessed independently, which we believe will make metadata more useful and accessible to readers.

More detailed functional design and implementation for Editor and Portal as ongoing work will be specified and deployed in the near future.

3 Discussion

The *"Article of the Future"* breaks through the traditional structure of papers. While the papers of Ontology-based platform are based on the structure features of *Cell,* they employ more semantic web techniques to each rhetorical chunk.

Table 1. Structural Comparison with Article of the Future and Traditional Article

Article / Features	**Traditional Article**	**Article of the Future**	**Article of Ontology-based Platform**
Paper structure	indivisible linear structure	divisible linear structure	divisible Ontology-based rhetorical structure

As illustrated in Table 1, we can tell the revolution in paper structure's evolution. In an ontology-based structured article, A*uthor* can provide metadata for every rhetorical chunk, and then an *Editor* may adopt the design methodology of supplemental semantic information to data and other materials that support or relate to main conclusions of each rhetorical chunk. But this supplemental information is considered as additional information or references. So divisible ontology-based structure can help readers explore papers conveniently and abundantly.

On semantic web, ontology plays a most important role. Because it strictly defines concepts and relations of such concepts, it can promote knowledge sharing and reusing in the knowledge level. Table 2 shows the article of ontology-based platform with features of introducing multiple types of data and semantic links to support the conclusions of the paper. In these papers, abundant data information (figures, audio resources, movie resources, descriptive text, etc) need to be constructed in the ontology to apply metadata describing these resources. What's more, the abundant ontology-based semantic search can generate semantic links and concepts, which provides multiple metadata to both whole paper and paper parts.

Table 2. Data Feature Comparison with Article of the Future and Traditional Article

Article Data Feature	**Traditional Article**	**Article of the Future**	**Article of Ontology-based Platform**
Figures and tables	main text (e.g. by pdf file)	main text & Individual supplemental information(by multiple formats file)	main text& Individual supplemental information (by multiple formats file, Semantic links, etc)
Experimental procedures	main text (e.g. by pdf file)	main text & Individual supplemental information (by pdf file, multimedia files)	main text & Individual supplemental information (by xml file, multimedia files)
References	Based on whole paper (e.g. by pdf)	Based on whole paper& each section (by pdf)	Based on whole paper& each section (by xml, Semantic entities and concepts links)
Readers' comments	None	Based on whole paper	Based on whole paper or each section
Multimedia files	None	Movies, Audio Clips, etc	Movies, Audio Clips, Semantic links, etc

Acknowledgement

This research is partially supported by China Scholarship Council (CSC). And many thanks to Prof. Fausto Giunchiglia for his inspiring suggestions.

References

[1] Thompson, S.A., Mann, W.C.: Rhetorical structure theory: A theory of text organization. Technical report, Information Science Institute (1987)

[2] de Waard, A., Tel, G.: The abcde format enabling semantic conference proceedings. In: SemWiki (2006)

[3] Groza, T., Handschuh, S., Mönller, K., Decker, S.: Salt – semantically annotated latex for scientific publications. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 518–532. Springer, Heidelberg (2007)

[4] Xu, H., Zhang, C.: Towards a pattern-based approach for scientific writing and publishing in Chinese. In: Chowdhury, G., Koo, C., Hunter, J. (eds.) ICADL 2010. LNCS, vol. 6102, pp. 264–265. Springer, Heidelberg (2010)

[5] Xu, H.: A semantic pattern approach to managing scientific publications. In: Kim, T.-h., Adeli, H. (eds.) AST/UCMA/ISA/ACN. LNCS, vol. 6059, pp. 431–434. Springer, Heidelberg (2010)

[6] Xu, H.: Managing ubiquitous scientific knowledge on semantic web. In: Kim, T.-h., Adeli, H. (eds.) AST/UCMA/ISA/ACN. LNCS, vol. 6059, pp. 421–430. Springer, Heidelberg (2010)

[7] Xu, H.: A pattern-based representation approach for online discourses. In: Kim, T.-h., Adeli, H. (eds.) AST/UCMA/ISA/ACN. LNCS, vol. 6059, pp. 378–384. Springer, Heidelberg (2010)

Multi-facade and Ubiquitous Web Navigation and Access through Embedded Semantics

Ahmet Soylu[1], Felix Mödritscher[2], and Patrick De Causmaecker[1]

[1] K.U. Leuven, Department of Computer Science, CODeS, iTec, Kortrijk, Belgium
{Ahmet.Soylu,Patrick.DeCausmaecker}@kuleuven-kortrijk.be
[2] Vienna University of Economics and Business,
Department of Information Systems, Vienna, Austria
felix.moedritscher@wu.ac.at

Abstract. Web content contains valuable information on the semantic structure of a site, which can be used to access and navigate the pages through ubiquitous computing environments. Semantic web approaches normally aim at modeling semantic relations and utilizing these models to provide enhanced functionality for humans or machines. In this paper we present an approach which focuses on using embedded semantics in order to achieve enhanced web access and navigation for the ubiquitous environments. Precisely we propose specifying and extracting microformat-based information within the web server and delivering it along the semantic structure of the site. We also describe our first prototype, the Semantic Web Component (SWC), and report on first experiences which evidence benefits in terms of less internet traffic and reducing the delivery of irrelevant information thus increasing the web accessibility as well as the navigability in ubiquitous environments.

Keywords: Ubiquitous Computing, Pervasive Computing, Embedded Semantics, Web of Data, Web Accessibility.

1 Introduction

The main motto of Ubiquitous Computing (UbiComp) [1, 2] deals with employing a variety of computing devices and applications, which are spread around the human environment, to seamlessly facilitate daily life through anytime and anywhere service and information access. These devices and applications need to communicate effectively so their behaviors and states can be synchronized (i.e. device/application interoperability). Furthermore, they need to share and understand available information to be able to deliver information and services relevant to users' context (i.e. data interoperability). The Web provides an appropriate framework respectively following two complementary approaches [3]: *(1) A communication-application space* which aims at enabling various mobile and stationary devices, including sensors and embedded devices, to get connected over the Internet. Consequently these devices can deliver their services to the each other and use available web applications and services while bringing them to the user environment (i.e. Web of Things [4]). *(2) An information space* which focuses on utilizing the Web as an ultimate information

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 272–289, 2010.

source [5], envisioning a web environment being one huge virtual, readable and writeable database rather than a document repository (i.e. Web of Data [6]). Consequently, the so-called ‘Semantic Web’ aims at increasing the utility and usability of the Web by utilizing semantic information on data and services [7].

In this paper, we will focus on the Web as an information space. The devices in UbiComp environments are expected to interact with each other, so that, they form a functional unit, i.e. virtually representing a computer. Hence, this computing network requires processable data to be readily available, however present web environments rather correspond a document repository [6]. The Semantic Web suggests a set of standards to overcome this problem so that machine readable data can be provided through the Web. XML, RDF, and OWL have been widely used for exchanging messages, modeling the application context [8], and describing services etc. Although each of these languages aims at different purposes at different levels (e.g. syntactic or semantic), the main problems can be summarized as follows. *(1) Redundancy of the information*: Web information can be presented in two distinct facades: *(a) human readable facade* and *(b) machine readable facade* of the information. Structurally separating these two facades requires information to be duplicated both in the form of HTML and in the form of RDF, XML etc. thereby causing synchronization and consistency problems. *(2) Loss of simplicity:* The main reason behind the success of the Web is its simplicity; anyone can use a basic text editor to create a web page. Hence, creating an RDF or XML document and uploading that external file dedicated to a machine-readable use remains forbiddingly complex [9], decreasing the accessibility. A complete picture of the Web's full potential should consider its human impact, as people are the most significant components [6].

A response to such considerations is embedded semantics [3] - eRDF, RDFa and microformats - which allow in place annotation of information without coding an external XML or RDF document and without duplicating the information. However such a solution imposes an extra burden, that is, extraction. Embedded information needs to be extracted out from (X)HTML. Although there exists a variety of client side applications, like the Firefox add-on ‘Operator’ (http://addons.mozilla.org/en-US/firefox/addon/4106) which detects and extracts embedded information, the restricted resources (i.e. limited memory and screen size, limited internet connection bandwidth, limited processing power, etc.) of mobile and embedded devices available through UbiComp environments make extraction of semantic information from web pages a non trivial task; in particular if pages include a high amount of multimedia content as well as textual, informational and structural elements.

The Web is supposed to be the main information source for UbiComp environments, hence it is important to ensure web accessibility through different devices with varying technical capabilities. In order to countervail the aforementioned critical issue, this paper presents our solution proposal called Semantic Web Component (SWC) for the server side (see Figure 1) and a basic query language, namely Web Query Language (WQL). SWC enables users and devices to access and navigate websites along their semantic structure. Thus, (human and non-human) actors can interact with related information only, not being confronted by irrelevant content. It reduces the size of information to be transferred and processed drastically and fosters the visualization of websites on devices with smaller displays, thereby providing increased accessibility and an efficient integration into the UbiComp environments. Moreover, WQL allows clients to submit basic queries through URL by HTTP GET method in order to retrieve only the content of interest while it is also used as a part of semantic navigation mechanism.

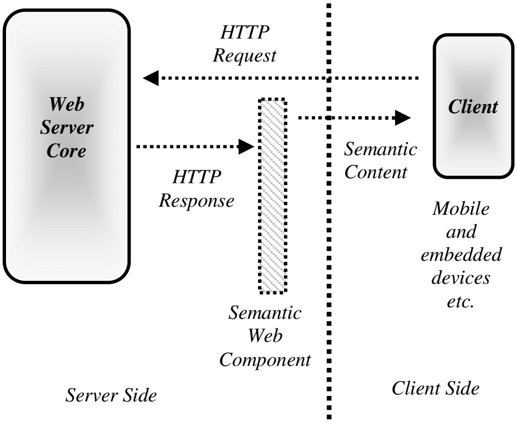


Fig. 1. The Semantic Web Component extracting the semantic content from the response

Since the component is an integral part of the web application server, any website hosted by such a server is covered by the functionalities of SWC. Such an approach is superior to the client side solutions in the sense of its uniformity and ease of employment by both the users and the machines. Furthermore since it imposes less responsibility to the client side, it might foster rapid employment of semantic web technologies and their standardization. The underlying technology utilized in SWC is built upon the embedded semantic technologies eRDF, RDFa and microformats. We also propose in this paper a description language for microformats to overcome its particular drawbacks (e.g. independence and extensibility) which in turn enables this technology to be compliant with our approach. Finally, the practical applicability of SWC is shown through the description of our first prototype implementation based on wide-spread concept of microformats and the description language.

The rest of the paper is structured as follows. In section 2, the role of embedded semantics in web technologies is briefly situated, and the language for microformats is described. The basic approach behind SWC with respect to current literature is explained in section 3. In section 4, the design and architecture of the SWC is presented. Section 5 evaluates our approach and discusses the related work. Finally section 6 concludes the paper and refers to our future work and its driving mantra.

2 Embedded Semantics

According to [10], the World Wide Web (WWW) is intended to be for humans while we believe that Semantic Web approaches rather lead to a more technologized Web, the ‘Web for machines’. Although these two facades of the Web coexist, unification in the structural sense is possible. However the Web for humans should not be compromised for the sake of the Web for machines. Embedded semantics imposes nearly no change in the current web technology and provides a simple and human-centered solution. Such an approach can be summarized by the four layers of information abstraction [2, 11]: *(1) storage layer* (e.g. tuples), (2) *exchange layer* (e.g. XML, JSON, RSS, etc.), (3) *conceptual layer* (e.g. OWL, RDF, etc.), and (4) *representation layer* (e.g. (XHTML, RDFa, eRDF, Microformats). The representation layer is not well studied, and the potential of the embedded semantics remains untouched to a large extent.

Human readable facade of the information through (X)HTML	Machine readable facade of the information through RDF	Uniform representation of the both facades through microformats
<div > Geo:	<rdf:RDF xmlns:rdf="…"	<div class="geo">Geo:
<span > Latitude :	xmlns:geo="…">	<span class="latitude">
30.386142	<geo:Point>	30.386142
</span>,	<geo:lat>30.386142	</span>,
<span> Longitude :	</geo:lat>	<span class="longitude">
120.092834	<geo:long> 120.092834	120.092834
</span>	</geo:long>	</span>
</div>	</geo:Point>	</div>
	</rdf:RDF>	

Fig. 2. Human readable facade, machine readable facade, and uniform facade of information

Embedded technologies use the attribute system of (X)HTML to annotate semantic information so that two facades of information are available in a single representation. Furthermore, employing the attribute system of (X)HTML allows developers to associate an external style sheet with the (X)HTML document to give any desired look and feel thereby loosely coupling the presentation and the information. In Figure 2, the first and the second code segments depict the human readable and machine readable facade of the information respectively. The last code segment demonstrates how these two facades can be combined into a single representation by means of embedded semantics. Embedded semantics approaches do not require altering current web technology and the approach itself is as easy as the Web itself since the required hand-on skills are modest. Apart from allowing machines to access machine readable information it also provides better user experience. For instance, users can export or copy some portion of the information from one web page to another one or an application by a single click [12]. In the followings we will elaborate on embedded semantic technologies.

Microformats: This community-driven approach provides a vocabulary and syntax to represent commonly known chunks of information such as events, people etc. Microformats use 'class', 'rel' and 'title' (X)HTML attributes to define domain specific syntaxes. It adopts well-known vocabularies such as vCard for hCard, iCal for hCal etc. Once its vocabulary and syntax are fixed, they should not be changed anymore. While Microformats can encode explicit information to aid machine readability, they do not address implicit knowledge representation, ontological analysis, or logical inference [13].

eRDF: eRDF also uses existing (X)HTML attributes 'class', 'rel' and 'title'. Unlike microformats it is based on the RDF framework that means it does not impose any pre-defined vocabulary. However it is not fully conformant to RDF.

RDFa: RDFa introduces new attributes 'about', 'resource', 'instanceof', 'property' and 'content'. These attributes are not yet supported by the current (X)HTML standard but expected to be included in the future. It is also based on the RDF framework and aims at reflecting the full capability of RDF.

While microformats are limited in flexibility, other techniques such as RDFa and eRDF provide more generic data embedding. Being based on RDF, eRDF and RDFa enable users to mix and use different name spaces. Microformats use a flat name space which is already predefined and cannot be extended or remixed. A microformat requires its own parser while generic parsers can be used with eRDF and RDFa. [14] lists four criteria for embedding semantic information. *(1) Independence and extensibility:* A publisher should not be forced to use a consensus approach, as she knows her requirements better. Web users originate from different communities, and thus follow their own local semantics for data interpretation and representation [12]. *(2) Don't repeat yourself (DRY):* (X)HTML should only include a single copy of the data. Hence modification can be done at one place which avoids consistency and synchronization problems. *(3) Locality:* When a user selects a portion of the rendered (X)HTML within his browser, she should be able to access the corresponding structured data (e.g. with a contextual menu). *(4) Self-containment:* It should be relatively easy to produce a (X)HTML fragment that is entirely self-contained with respect to the structured data it expresses.

Accordingly, an evaluation of technologies [14] is given in Table 1. eRDF has to provide vocabulary related information in the (X)HTML head while microformats either assume clients to be aware of all available syntaxes beforehand or require a profile URI to be provided for extraction. Microformats and eRDF lack self-containment because it is not possible to re-use eRDF or microformat information without requiring vocabulary specific information. On the other hand microformats lack of independence and extensibility since they are based on pre-defined vocabularies and they require a community consensus.

Table 1. An evaluation of embedded semantics technologies based on four main criteria

Criteria/ Technology	Independence and Ex.	DRY	Locality	Self-containment
RDFa	Yes	Yes	Yes	Yes
eRDF	Yes	Yes	Yes	Not fully
Microformats	No	Yes	Yes	Not fully

Although RDFa provides a far better solution in the technical sense, employment rates of these technologies do not seem to be in line with their technical merits. A recent estimate shows that microformats are used in hundreds of millions of web pages (http://microformats.org/blog/2007/06/21/microformatsorg-turns-2/) while deployment of eRDF and RDFa still remains weak. This is mainly because of its simplicity which might be the fifth criteria. On the client site, the user interaction paradigm is switching from passively consuming content (i.e. surfing on the Web) to actively contributing (i.e. authoring/editing information on the Web) via weblogs, wikis, and user driven contents in general [12]. Therefore, having users as active contributors of the Web (e.g. as content authors, consumers and even as application developers) increases the demand for the simplicity. Particularly, microformats offer an easy mechanism for humans to publish information, and it lowers barriers for publishers by following a publisher-centric solution rather than a parser-centric one (see http://microformats.org/wiki/principles). These merits are due to the basic principles on which the research, design and development of microformats are based.

Since adoption of microformats is wide, we implemented our first prototype of SWC based on microformats. However it is important to note that eRDF and RDFa are compliant with overall idea and to be covered by SWC. However, lack of independence & extensibility and self-containment of microformats are important barriers for the SWC. This is because it is not possible to define custom microformats, and the consumer (i.e. SWC) is expected to have a pre-knowledge of the syntax and vocabulary of available microformats. Methodologically, there are two possibilities to describe microformat-based semantics embedded within the content. (1) Providing a XSL schema and, thus, describing what content chunks should be extracted in which particular way, as also done with GRDDL ('Gleaning Resource Descriptions from Dialect Languages') transformations [15]. (2) Developing an own simplified and generic way to specify embedded semantics. In this paper, we decided to follow the later approach. The first approach imposes dependence to the client by assuming that it supports XSL, and more importantly the client has to accept the extracted information in whatever form it is extracted to by XSL. We want to reduce complexity of the semantic description language and try to avoid describing how to extract the microformat-based semantics. However, a mapping from our data model to XSL can be easily made up. Such a description language provides a further layer of abstraction at the transformation side. Firstly, it allows custom microformats to be defined, thereby alleviating independence and extensibility problem. Secondly it allows clients to understand the exact structure of the available microformat rather than assuming that the client is already aware of all possible microformat syntaxes and vocabularies. Although this approach still does not fully satisfies self-containment, since existence of vocabulary and syntax specific information is required, it is superior to client pre-knowledge approach. Furthermore it allows the client to extract available information in any way to any form without imposing any technological dependence.

Allsopp lists 12 concrete examples of microformat specifications, beginning from elemental ones, like rel-license, rel-tag or VoteLinks, up to compound microformats, such as hCard, hCalendar or hAtom [16]. In practice there exist even more specifications (see www.microformats.org). In order to describe all these embedded semantics, a data model has to consider the following aspects in terms of required or optional attribute fields. *(1) Type:* The first issue to determine is if one wants to use an elemental or a compound microformat. *(2) Identifier:* Second, specifying embedded semantic requires, like all other resource description standards, some kind of identifier, so that applications or humans can differentiate between the semantic elements. *(3) Design pattern:* Third and mostly important, it is necessary to describe which design pattern one wants to address. Microformat design patterns comprise a formalism to 'reuse pieces of code that are generally useful in developing new microformats' [16]. In other words, design patterns determine which (X)HTML elements and attributes are used to define a certain microformat. Thus, we propose to describe such a design pattern according to these two entities: (a) the element name, and (b) the attribute name. Assuming that a microformat is always based on an attribute, we consider the element as optional and the attribute as required. Furthermore, it should be possible to combine elements according to different attributes. *(4) Label:* A user understandable label which can be used while representing the extracted information, so different parts of the information can be identified by the users. Albeit not mandatory, it has an absolute use for SWC (see section 5). *(5) Matching string:* In order to restrict the (X)HTML attribute of the design pattern, an optional field for string matching is introduced. Values of the specified (X)HTML attributes are evaluated on basis of string equivalents as well as regular expressions. *(6) Scope:* The scope, again, is optional and restricts the scope of the semantics within the web-based content. If given, the embedded semantic is valid within all DOM elements specified by this field. *(7) Selector:* Another optional field, the so-called selector, is necessary to define from which source ((X)HTML element text or attribute) the semantics has to be extracted. If no selector is specified the value of the element is used. Otherwise, an application might retrieve the value of the specified selector which, for instance, could be the title attribute. *(8) Reference:* The optional reference field is of use to refer to another, existing microformat. Such a mechanism is useful for compound microformats, i.e. to include elemental microformats. If referring to another microformat, all other fields except the identifier are ignored. *(9) Optional:* Finally, the optional field indicates that an elemental microformat is optional within a compound one, which means that this element is not required to detect the compound microformat.

```
1   <microformats>
2     <elemental id="xfn_met" label="XFN: People I met in person" pattern="rel" match="friend met" select="text" />
3     <elemental id="vote" label="VoteLinks: Vote for me! " pattern="a:rev" match="vote-*" select="title" />
4     <elemental id="vote_link" label="Click here to vote" pattern="a:rev" match="vote-*" select="href" />
5     <elemental id="all_links" label="All external links" pattern="a:*" match="http://* " select="html/body" />
6     <elemental id="fn" label="hCard: Get full name" pattern="class" match="^fn  | fn | fn$" select="text" />
7     <elemental id="url" label="hCard: Two variants for URLs" pattern="a:rel|div:rel" match="url" select="href" />
8     <compound id="vevent" label="hCalendar: exemplary event" pattern="class" match="vevent">
9       <elemental id="vevurl" label="Event URL" ref="url" />
10      <elemental id="vevsummary"  label="Event summary" pattern="class" match=”summary” select="text" />
11      <elemental id="vevstart" label="Start date" pattern="class" match="dtstart" select="title" />
12      <elemental id="vevend" label="End date" pattern="class" match="dtend" select="title" optional="true" />
13    </compound>
14    <elemental id="goal" label="AdeLE’s learning goals" pattern="adele" match="to *" scope="/html/body/content"  />
15  </microformats>
```

Fig. 3. An example XML binding for the proposed microformat description language is given

Figure 3 shows a possible description language for microformat-based semantics. In this example, seven elemental microformats (lines 2 to 7 and line 14) and one compound one (line 8 to 13) are specified. The first elemental microformat, namely 'xfn_met' (line 2), stands for a particular type of the commonly-known (X)HTML Friends Network (XFN) specification which can be determined with the rel-attribute having the value 'friend met'. For extracting semantics from such elements, the text-field (the value between opening and closing tag) has to be used via the selector-field. If no selector is given, an information extractor simply uses the value of the specified pattern. Having such a specification of an elemental microformat, any application, even browser plug-ins can detect and extract this kind of semantic information from web-based content if supporting our data model.

3 The Semantic Web Component

In this section the basic idea behind the SWC is introduced. The driving challenges are twofold; *extraction* and *unification*. Firstly, the annotated information needs to be extracted out of the original content. Although the client side approaches are currently common, we will argue for a server side approach. Secondly, the fact that different embedded semantics technologies are available requires unifying the use of these technologies. We advocate this diversity and move unification to the sever side extraction mechanisms rather than opting for a single technology. We start with referring to the related literature with respect to these two points, so that the idea and our understanding can be situated on a concrete grounding. The literature provided in this section is not exhaustive, yet we have selected particular works characterizing our basic challenges.

In [2], the possible benefits of embedded semantics for the UbiComp environments are elaborated. Authors identify information as one of the important elements of their upper context conceptualization and state that embedded semantics is useful for representing different contextual characteristics of the information so that such contextually annotated information can be delivered to the users in an adaptive manner. Furthermore, they describe a web service which extracts and collects embedded information for learning resources from web pages. The harvested information is stored in a semantic database, allowing other clients to query its knowledge base through SPARQL queries. A similar approach has been employed in [17]. In the scope of earth observation [18] uses RDFa for identifying embedded information through a browser extension [19]. The information extracted is either used to populate ontologies with the extracted information or to be stored in the semantic repository. [2, 17, 18, 19] show that there exist different ways of harvesting embedded information. On the one hand, client side tools, such as Operator or Semantic Turkey, are used to extract or distinguish the annotated information from web content. The main drawback of these approaches is that they require a client side mechanism to extract information, so computing resources of the clients are used. Furthermore the whole content has to be downloaded to the target machine which is problematic due to the network load. On the other hand, third party web applications or services, as demonstrated in [2, 17], are utilized. In this case, the semantic search services provided usually duplicate the information by means of storing extracted information which is against one the driving principles of embedded semantics, namely the DRY principle. Furthermore, it imposes a dependency to other third party web applications or services. Clearly such approaches are not feasible for the

UbiComp environments since they are expected to include many small devices having low-bandwidth. Considering unification matter, in [14] proposes a mechanism, namely hGRDDL, to transform microformat embedded (X)HTML into its RDFa equivalent. This mechanism aims at allowing RDFa developers to leverage their existing microformat deployments. They advocate that such a solution can allow RDFa to be a unifying syntax for all the client-side tools. There are two important problems in this approach. First of all, developers need to provide vocabulary and syntax for each microformat to be transformed. Such a problem can be solved by using the description language which we have proposed in the previous section. However we disagree with unification by means of a unified syntax, indeed a technology not only a syntax, since a decision between microformats and RDFa is a tradeoff between simplicity (i.e. usability) and functionality.

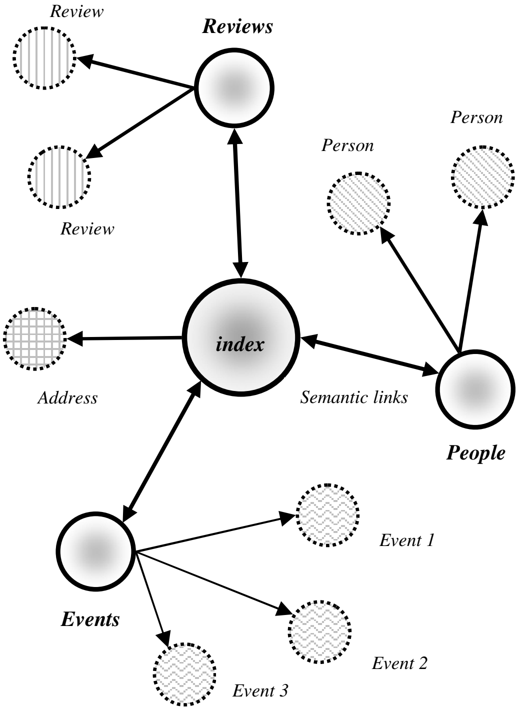


Fig. 4. Semantic information network (map) referring to semantic structure of a web site

Therefore we propose a solution proposal called Semantic Web Component (SWC) for web application servers. Since the Semantic Web is an important construct of the tomorrow's ubiquitous Web, application servers should be able to deliver two facades of the information directly and should allow both humans and machines to interact and navigate through them. SWC resides in the server side and observes the requests and the responses between the client and the server. When a client requests the machine readable facade of the information, instead of returning all the (X)HTML content it filters out only semantically annotated information. One option is that the information is extracted and represented in a (X)HTML form (i.e. reduced (X)HTML content). All other information, which is not annotated, is simply discarded. Such an approach treats the pages of a website as a set of nodes where each node might contain instances of several types of embedded information. Embedded information can be elemental, including only one single and independent chunk of information, or compound consisting of at least two elemental embedded information. Each node (i.e. page) also has links to other pages having embedded information. We call these links *semantic links*. The approach, we named it *semantic information network* (or map), is visualized in Figure 4. This facade is still the human facade, however it represents reduced content and allows user to navigate through the semantic information network of a website. On the other hand, if the machine asks only for machine readable facade, the component converts extracted information to the XML or RDF.

We summarize advantages of such an approach with respect to the aforementioned works in the followings. *(1) Direct and seamless access* to different facades of the information without imposing any burden to the client side, e.g. no need for data extraction. *(2) Enhanced user experience:* users are usually lost in the abundant information space of the Web where valuable information is hidden in the information sea and presentational and structural elements. Users can simply access the information they do require. *(3) Increased accessibility:* mobile and embedded devices in the UbiComp environments can use both facades of the information. (X)HTML representation of the reduced information will enable them to deliver web information to anyplace while machine readable form of the information will enable devices to process and use the web information. *(4) Higher network efficiency:* the device do not need to retrieve all the (X)HTML content from the server, hence the amount of information travelling in the network decreases. *(5) Centralized solution:* it does not impose use of a common syntax, technology or dependency to any other service; everything is unified at the server side.

We introduce three particular scenarios to demonstrate the use and benefits of such an approach.

Scenario-1: A website of a cinema company provides recommendations for the movies of the season. The site consists of following pages: ‘Events’, ‘People’, and ‘Reviews’. Each movie is considered as an event in the ‘Events’ page. ‘Reviews’ page includes the reviews about the movies and each review is provided by a registered reviewer. The ‘People’ page contains information about the registered reviewers. The information on these pages is annotated by using and hCal, hReview and hCard microformats respectively. Accordingly the semantic information map of the website is shown in Figure 4. A user wants to see a movie tonight. He does not have much time to surf through the website to find a proper movie. Furthermore, he only has his mobile phone around which has internet access. However his mobile device’s connection and screen properties are at a low level. Since the website is hosted by a server which has SWC enabled, the user simply sends a request through his mobile phone. His browser implicitly tells the server that it only requests annotated information. Server returns to the user the list of semantic information available in the index page; people, events, and reviews. The user selects the reviews option to see the movie reviews, and the server returns the list of available instances which are identified with the titles of the movies. The user selects a movie in which he might be interested and reads the reviews. He really likes one of the reviews and wants to see who wrote it to be sure that he can trust the quality of this information. He navigates back to the first page and repeats same procedure to see the details of the reviewer. The user decides that the movie is worth to go and the reviewer is really appropriate. Then he goes through the events page to see the schedule.

Scenario-2: Based on the previous scenario, the company wants to place small terminals to some particular places through its main hall. These terminals are expected to provide basic information available in their Web page; events (i.e. movies), reviews (i.e. movie reviews) and people (i.e. reviewers). However the budget of the company is limited to buy only low cost devices which only have text-based presentation capabilities, besides this is the only desired functionality of the company. These devices are connected to the cinema's website through normal internet connection. The browser implicitly tells server that it requires semantically annotated information; hence these devices provide the same navigational mechanism as described in the first scenario.

Scenario-3: A recommender system suggests activities to users; therefore it has access to their agendas and profiles. To find appropriate activities, the recommender system harvests the embedded information from various websites in the same way mentioned in the first scenario. It uses ontological reasoning and has a terminological base providing the upper and domain ontologies as well as a knowledge base with the ontology instances. The terminological base is already pre-defined while instances are collected on the fly through harvesting semantic information from different websites. Since the harvested websites are supported by SWC, the recommender agent does not need to be aware of different embedded technologies. The machine readable facade of the information is directly sent from the servers. The recommender system reasons that the user has nothing scheduled on Saturday night and she is keen on horror movies. The agent submits WQL queries to various entertainment sites asking for events scheduled for Saturday night. According the information harvested from the cinema's website, it finds out that there is a new horror movie which is highly ranked on this Saturday night. Therefore this movie is recommended to the user. Obviously there are many other possibilities for recommendations, like a ranked list of the events or the most topical reviews.

4 Design and Implementation

We have set multi-facade and ubiquitous web navigation and access into practice by implementing a prototypic SWC component which includes the microformat descriptors, i.e. the microformat descriptions of the embedded semantics to address, and the semantic extraction functionality. Figure 5 demonstrates the architecture of our component which is realized in the form of Apache modules. Thereby a module is a self-contained plug-in which may implement core functionalities, a general purpose service, a small but vital function or a single purpose application of the Apache Web Server [20].

The SWC component is composed of two modules, namely *mod_semantic* (i.e. the handler, which provides the functions to be performed when URL requests are sent to the server – see http://httpd.apache.org/docs/2.0/handler.html) and *mod_grddl* (i.e. the filter, which processes the data sent or received by the server – see http://httpd.apache.org/docs/2.0/filter.html). The former module implemented in C listens to the client requests and provides the appropriate functionality. The latter module is responsible for information extraction. The client is expected to send a contextual header value named 'is_semantic' which maps to three distinct modes. The current implementation is based on PHP, for simplicity, thus it works more like a proxy. The modes and the corresponding actions are described in the following.

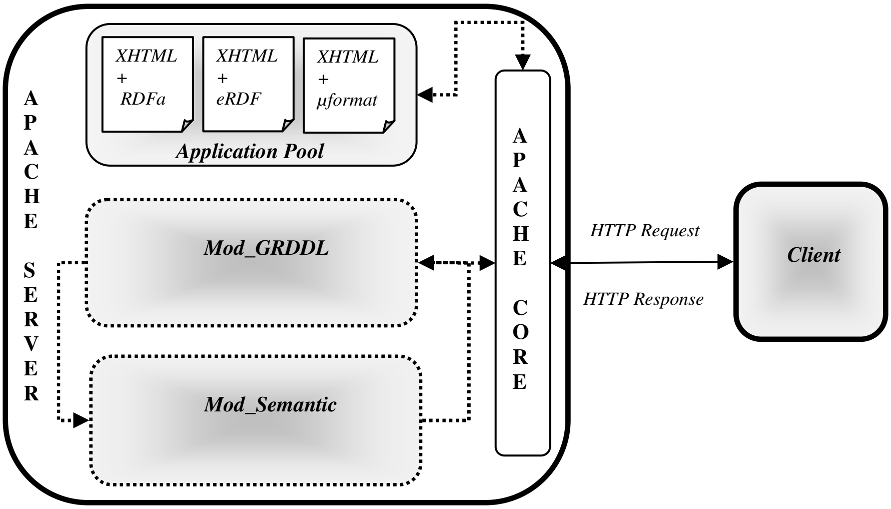


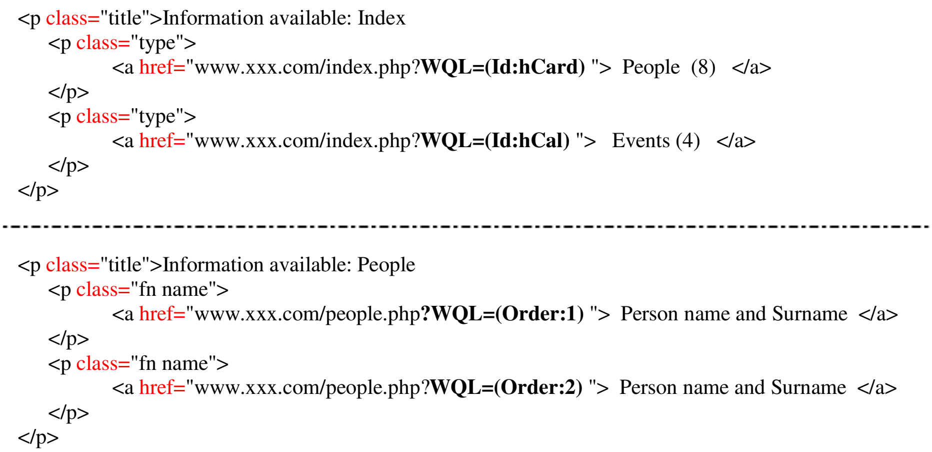
Fig. 5. Overall architecture of the component within the Apache Web Server is depicted

Case-1 (Mixed facade mode): If the value of the contextual header element is equal to '0', the handler module calls the requested resource and delivers it as it is. This case does not change original behavior of the Web server. The returned (X)HTML document also includes embedded information, therefore this mode is called as mixed facade mode.

Case-2 (Machine facade mode): If the value of the contextual header element is '1' then the handler module calls the original resource, and forwards it to the filter module (e.g. after being generated or for dynamic content such as a PHP script). The filter module extracts and delivers embedded information in as RDF or XML and according to the descriptors of the semantics available within the source document.

Case-3 (Human facade mode): If the value of the contextual header element is equal to the '2', the handler module calls the original resource and forwards it to the filter module. The filter module retrieves the descriptor within the source (X)HTML document, extracts the embedded information in the form of (X)HTML and forwards the response. This approach enables user to move inside the semantic (X)HTML structure, i.e. semantic information map previously shown in Figure 4. We also call (X)HTML snippet returned to be the reduced (X)HTML content since it is still represented in (X)HTML but the non-annotated parts of the document are discarded.

In the last mode (i.e. case-3), the content retrieval is iterative. For instance, if a user initiates navigating a website through a page EP (*entry page* or *entry node*: the first accessed page, which does not necessarily need to be the index page) the component returns a set of available types of embedded information and the number of instances available for each type in the EP. The returned message will be a simple (X)HTML response which presents human understandable labels of extracted microformats which are available through the label attributes of the descriptor. If the user selects a type of embedded information, the SWC returns the list of available instances together with a small description, e.g. title of each instance. The user can further select an instance; then the information available for this instance is returned as a response. If the instance is one elemental embedded information or a set of several elemental embedded information the whole content is displayed. However if the embedded information includes another compound embedded information the user needs to further drill down. Up to this point all the navigation within the semantic information was available in the EP. However users can also navigate into the other pages, which is only possible by annotating links which reference to the other pages involving semantic information and omitting the non-annotated links. In order to achieve annotated links, SWC uses the 'nav_link' attribute value which also can be seen as a type of embedded semantics to annotate links between the pages. Rather than fetching the whole content of the page or all the available semantic information through the page, this approach allows user or machine to partially retrieve semantic information. The same iterative procedure can also be applied for the machine facade mode if required.



```
<p class="title">Information available: Index
    <p class="type">
            <a href="www.xxx.com/index.php?WQL=(Id:hCard) "> People (8) </a>
    </p>
    <p class="type">
            <a href="www.xxx.com/index.php?WQL=(Id:hCal) "> Events (4) </a>
    </p>
</p>
------------------------------------------------------------------------------
<p class="title">Information available: People
    <p class="fn name">
            <a href="www.xxx.com/people.php?WQL=(Order:1) "> Person name and Surname </a>
    </p>
    <p class="fn name">
            <a href="www.xxx.com/people.php?WQL=(Order:2) "> Person name and Surname </a>
    </p>
</p>
```

Fig. 6. Basic (X)HTML templates for blank and parametric calls are depicted respectively

In order to realize such a navigational procedure, we have introduced a simple *presentational template* and a query language named *Web Query Language* (WQL). The basic presentational template is composed of '<p>' and '<a>' elements. It is used to only deliver the information at the most basic level. The class attribute system is still utilized to annotate the information, so any desired look and feel can be given to the information retrieved by means of CSS. In Figure 6 an example is depicted. The first part of the figure belongs to the initial call of the entry page. Hence detected types of embedded information and the number of the instances available are listed. The second part of the figure represents a call to a particular type of embedded information in the corresponding page. This time available instances and their descriptive titles are displayed. The navigation is done by using HTTP GET through the URLs available in the basic presentation. Each get method submits a basic WQL query which either includes the unique id (i.e. the type) of the embedded information or the order number of the instance. So a user can navigate into a specific type of embedded information or to a specific instance. When there is no WQL query parameter given in the URL, we name it a *blank call* (e.g. initial call to the entry page). Conversely, if some WQL parameters are defined, the call is named as *parametric call*. The motivation for and the details of WQL are explained in the followings.

The proposed query language arose from the need to identify the type and instances of embedded information to be navigated, e.g. to enable a user to see the instances of people available at the cinema website or to select one specific person. The basic structure of a WQL query is as follows:

[URL]?WQL=([conditions])

The WQL query is provided right after the requested URL with a reserved GET parameter 'WQL'. The conditions have to be specified inside the bracket symbols '(' and ')'. The constructs of the query language are listed and explained in Figure 7.

Symbols	Description
:	Usage: *<elementName> : <characterString>* This symbol works as "LIKE" operator in SQL. It is used to describe if a pattern matches given character string. Example WQL query: *WQL=(fn_name:”Tommy”)*
;	Usage: *<firstCondition> ; <SecondCondition>* This symbol corresponds to AND operator in SQL. Example WQL query: *WQL=(fn_name:”tommy”;fn_surname:”Brown”)*
,	Usage: <condition_1> , <condition_2> This symbol corresponds to OR operator in SQL. Example WQL query: *WQL=(fn_name:”Tommy”,fn_name:”Alice”)*
Reserved variables	Description
Order	Usage: *Order: <positiveInteger>* This reserved variable behaves as a predefined unique key which spans all the instances of different types of embedded information available through a single page. Example WQL query: *WQL=(Order:2)*
Id	Usage: *Id: <string>* This reserved variable is used for accessing a specific type of embedded information. Example WQL query: *WQL=(Id:hCard)*

Fig. 7. The basic constructs of the Web Query Language (WQL) is given

WQL realizes the most basic facilities of a SQL like query language. We have introduced three symbols which represent the 'LIKE', 'AND', and 'OR' operators of SQL. When the ':' operator is used, the wildcard character '%' is automatically added to the both sides of the string, which indicates a string of any length. However when applying it with reserved variables, it works in the same way as the equality operator in order to prevent any ambiguity (e.g. Id:1 vs. Id:12). The element name used with the ':' operator can refer to any of the words in the vocabulary and to any type of the embedded information available. We have also introduced two reserved variables which are 'Id' and 'Order'. The 'Id' variable is used to match with the 'id' attribute in the microformat description language to identify the type of embedded information thereby allowing a particular type of embedded semantics to be retrieved. The 'Order' variable is used to access a specific instance by pointing to a unique number for each instance. This unique value is derived based on the assumption that the order of each instance occurring in a page remains constant, and it is in increasing order depending on the place of the instance. Only reserved variables start with an upper case letter, and the client can provide as many variable-value pairs desired within a WQL query to be matched with the vocabularies of the embedded information. Although the current implementation of the WQL is mainly for navigational purposes, it will be further developed and validated, so it can be used for querying Web pages through URLs at the most basic level as demonstrated in Figure 7. Therefore WQL is intended to be simple. Since WQL is less expressive than SPARQL underlying implementation can be realized through mapping WQL queries to SPARQL queries in order to have a standardized implementation. Search/Retrieval Using URL (SRU - http://www.loc.gov/standards/sru/) and Yahoo! Query Language (YQL - http://developer.yahoo.com/yql/) are similar approaches to WQL. The former focuses on XML and the latter is proprietary since every query is executed through their central API. WQL approach assumes any server to be able to execute WQL queries through SWC. In terms of expressivity, WQL is less expressive than YQL and SRU, however this is because we opt for simplicity at this point. Further extensions to WQL is mainly intended to follow SRU since it is based on a similar simple syntax while YQL follows a more complex SQL/SPARQL like syntax.

5 Evaluation and Discussion

A preliminary evaluation of the component and the description language has been done for a website containing real-world data about a research group. The website includes two types of embedded information, precisely people (i.e. research members) and events (e.g. seminars), comprising 26 instances of people and 15 instances of events. These instances are embedded in 31 pages of the website. The result of a blank request is visualized on the left-hand side of Figure 8, while the result of the request for all people instances is shown on the right-hand side.

The most basic evaluation of our approach can be done from an UbiComp perspective in twofold: *(1) network traffic:* comparison of the amount of information downloaded in the mixed mode and the amount of information downloaded in the human facade mode, *(2) network calls:* comparison of the amount of page requests while user is navigating in mixed mode and the amount of the page requests while user is navigating in the human facade mode. The measurements are based on the fact that all the available semantic information instances need to be retrieved in both facades of the navigation and within one single session for each of the facades.

This case study comparing the human face mode to the mixed mode (full web application) of SWC evidences one important benefit of our approach: On the one hand, the difference between the amount of information transferred during mixed facade navigation and the amount of the information transferred in human-facade mode is drastic. Due to the fact that a page with all its basic presentational markups contains typically around 27 KB of data in average (without considering the multimedia content!), in the first mode a total amount of 849.2 KB data is downloaded. In the second mode, however, the data transferred is reduced to around 110 KB, as a chunk of embedded information has a size of 1-2 KB in average. On the other hand, the number of network calls done in the two sessions increases from 31 calls in the first facade to 51 for the second facade. The difference depends on the structure of the website and structure of the embedded information available. Overall, the increase in the amount of network calls seems admissible since the amount of information downloaded in each call is considerably small. The significant reduction of transferred data clearly favors our component.

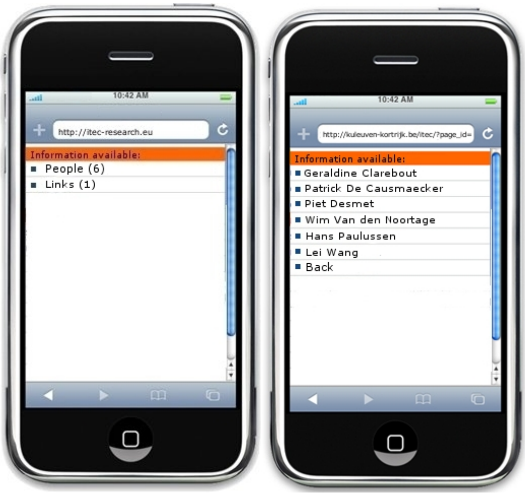


Fig. 8. Blank and parametric requests to the research group web site are shown respectively

Due to the navigation along the semantic structure of a website, we see clear advantages for mobile devices and web accessibility. One the one hand, less data is transferred to the web client. On the other hand, irrelevant information is filtered out. The second issue however could be problematic for paid advertisements. In the literature there are several studies addressing web access through mobile devices with limited sources. Amongst others, [21] reports on website personalizers observing the browsing behavior of website visitors and automatically adapting pages to the users. Moreover [22] examine methods to summarize websites for handheld devices. In [23] authors employ ontologies (OWL-S) and web services in order to realize context-aware content adaptation for mobile devices. All these approaches either require authoring efforts, e.g. for creating ontologies, or are based on AI-based techniques which cost a considerable amount of computational processing. Our approach on the other side builds upon a simple specification of semantics embedded on a website and low processing efforts done by the web server. Anyhow the SWC enables users to access and navigate web content along their semantic structure thus reducing the traffic and providing personalized chunks of information, even for mobile devices.

6 Conclusions and Future Work

In this paper we argued for using embedded semantics (i.e. microformats) for UbiComp. Instead of building upon ontologies or complex mining techniques we proposed a description language for microformat-based information which is used by two Apache web server modules (the SWC) to enable users to access and navigate web content along the semantic structure of a website. In a first evaluation study we evidenced that SWC can reduce internet traffic as well as irrelevant content thus increasing its applicability for UbiComp environments and mobile devices.

The work presented in this paper serves as a proof of the concept, indicating the advantage of the overall approach. Accordingly, our future work involves exhaustive validation of the whole approach with a particular focus on usability. E-learning is one of our immediate application domain since our main research challenge is enabling adaptive ubiquitous learning, requiring us to ensure the accessibility of the web-based learning environments. Finally, we envision going beyond semantic and ubiquitous web navigation and extending our approach with respect to user interactions and user-driven development of web environments.

Acknowledgments. This paper is based on research funded by the Industrial Research Fund (IOF) and conducted within the IOF Knowledge platform ‘Harnessing collective intelligence in order to make e-learning environments adaptive’ (IOF KP/07/006). Partially, it is also funded by the European Community's 7th Framework Programme (IST-FP7) under grant agreement no 231396 (ROLE project).

References

1. Satyanarayanan, M.: Pervasive computing: vision and challenges. Pers. Commun. IEEE 8, 10–17 (2001)
2. Soylu, A., De Causmaecker, P., Desmet, P.: Context and Adaptivity in Pervasive Computing Environments: Links with Software Engineering and Ontological Engineering. J. Softw. 4, 992–1013 (2009)
3. Soylu, A., De Causmaecker, P., Wild, F.: Ubiquitous Web for Ubiquitous Computing Environments: The Role of Embedded Semantics. J. Mob. Multimed. 6, 26–48 (2010)
4. Dillon, T., Talevski, A., Potdar, V., Chang, E.: Web of Things as a Framework for Ubiquitous Intelligence and Computing. In: Zhang, D., Portmann, M., Tan, A.-H., Indulska, J. (eds.) UIC 2009. LNCS, vol. 5585, pp. 1–10. Springer, Heidelberg (2009)
5. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Sci. Am. 284, 34–43 (2001)
6. Ayers, D.: The Shortest Path to the Future Web. Internet Comput. 10, 76–79 (2006)
7. Mödritscher, F.: Semantic Lifecycles: Modelling, Application, Authoring, Mining, and Evaluation of Meaningful Data. Intl. J. of Knowl. Web Intell. 1, 110–124 (2009)

8. Perttunen, M., Riekki, J., Lassila, O.: Context Representation and Reasoning in Pervasive Computing: a Review. Intl. J. Multimed. Ubiquitous Eng. 4, 1–28 (2009)
9. Khare, R.: Microformats: the next (small) thing on the semantic Web? Internet Comput. 10, 68–75 (2006)
10. Huang, W., Webster, D.: Enabling context-aware agents to understand semantic resources on the www and the semantic web. In: IEEE/WIC/ACM Conf. on Web Intelligence (WI 2004), Beijing (2004)
11. Reichle, R., Wagner, M., Khan, M.U., Geihs, K., Lorenzo, L., Valla, M., Fra, C., Paspallis, N., Papadopoulos, G.A.: A comprehensive context modelling framework for pervasive computing systems. In: 8th IFIP Intl. Conf. on Distributed Applications and Interoperable Systems, Oslo, Norway (2008)
12. Mrissa, M., Al-Jabari, M., Thiran, P.: Using microformats to personalize web experience. In: 7th Intl. Workshop on Web-Oriented Software Technologies IWWOST 2008 (2008)
13. Khare, R., Çelik, T.: Microformats: A pragmatic path to the Semantic Web. In: 15th Intl. World Wide Web Conf., Edinburgh, pp. 865–866 (2006)
14. Adida, B.: hGRDDL: Bridging micorformats and RDFa. J. Web Semant. 6, 61–69 (2008)
15. Connolly, D.: Gleaning Resource Descriptions from Dialects of Languages. W3C (2004), `http://www.w3.org/2004/01/rdxh/spec`
16. Allsopp, J.: Microformats: Empowering Your Markup for Web 2.0. FriendsofED, Berkeley (2007)
17. Sabucedo, L.A., Rifón, L.A.: A Microformat Based Approach For Crawling And Locating Services In The Egovernment Domain. In: The 24th Intl. Symposium on Computer and Information Sciences, pp. 111–116. IEEE Press, Guzelyurt (2009)
18. Fallucchi, F., Pazienza, M.T., Scarpato, N., Stellato, A., Fusco, L., Guidetti, V.: Semantic Bookmarking and Search in the Earth Observation Domain. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part III. LNCS (LNAI), vol. 5179, pp. 260–268. Springer, Heidelberg (2008)
19. Griesi, D., Pazienza, M.T., Stellato, A.: Semantic Turkey - a Semantic Bookmarking tool (System Description). In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 779–788. Springer, Heidelberg (2007)
20. Kew, N.: Apache Modules Book: Application Development with Apache. Prentice Hall, Englewood Cliffs (2007)
21. Anderson, C.R., Domingos, P., Weld, D.S.: Personalizing Web Sites for Mobile Users. In: The 10th Intl. World Wide Web Conf., pp. 565–575. ACM, Hong Kong (2001)
22. Buyukkokten, O., Garcia-Molina, H., Paepcke, A.: Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices. In: The 10th Intl. World Wide Web Conf., pp. 652–662. ACM, Hong Kong (2001)
23. Forte, M., de Souza, W.L., do Prado, A.F.: Using Ontologies and Web Services for Content Adaptation in Ubiquitous Computing. J. Syst. Softw. 81, 368–381 (2008)

New Format and Framework for Managing Scientific Knowledge

Hao Xu[1,2,3]

[1] College of Computer Science and Technology, Jilin University,
Qianjin Street 2699, Changchun 130012, China
[2] Center for Computer Fundamental Education, Jilin University,
Qianjin Street 2699, Changchun 130012, China
[3] Department of Information Engineering and Computer Science,
University of Trento, Via Sommarive 14, Povo 38123, Italy
hao@disi.unitn.it

Abstract. Only few models or frameworks of making situated use of advanced technology in scientific publication scenarios are available. Moreover, most of the existing prototypes and applications lack specific expertises and semantics. The approach we proposed in our project is based on some fundamental theories especially specified for managing scientific knowledge. In this paper, we concisely address our conceptual model and methodology defined for ubiquitous scientific publications, along with the pattern approach to systematically and semantically solve the problems in the current age of scientific publishing.

1 Introduction

Up to date, a prestigious publisher always provides a highly recognizable format and presentation for its published papers. However, most of them haven't been largely changed during several decades. In the mid 1990's, the advent of Internet brought amazing opportunities for scientific journals. Online publishing thoroughly revolutionized the searchability and information discovery, tremendously increased the breadth and ease of access, and gradually allowed for dissemination of supplemental materials such as large data sets, comments and some related citation links online which could not be captured in traditional printed publications. But few have tackled the problem of how best to bring the magics of the new ICT technologies, especially of Web 2.0 and Semantic Web technologies, to bear on the structure representation, organization and presentation of the article itself. Thus, for most publishers the online paper of today remains essentially an electronic facsimile of the traditional print publication. The Journal of Cell[1] has taken a successful attempt to promote the direction of "Article of Future"[2],

[1] Cell: http://beta.cell.com/

[2] Article of the Future:
http://beta.cell.com/index.php/2010/01/
cell-launches-article-of-the-future-format

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 290–293, 2010.

but it is restrained by its respectively narrow discipline and types of literature, which is difficult to be applied to all kinds of scientific publications for more general potential readers anyhow.

The initial motivation of our project comes from a narrative of writing a PhD qualifying paper. To start with, the student uses Google Scholar[3], Citeseer[4] and DBLP[5] to accumulate his background knowledge to achieve the state of the art and generates a tentative "Gas" idea. Whereafter, he discusses it with his supervisor and colleagues face to face or via emails,meanwhile he attends interesting seminars, courses, related workshops and conferences, beginning to draft his "Liquid" paper. After several iterations, he finishes organizing and writing the qualifying paper by LaTex and then sends the "Solid" PDF file to the committee. He gets feedbacks from reviewers and checks the review forms item by item according to his paper to make a final modification.

Although some technical progress has been made in such scenario, at least several obstacles must be overcome before a semantic framework can be realized. Firstly, how to write a PhD qualifying paper, essentially what the structure of a qualifying paper should be followed and how to prepare both background knowledge and writing skills for it are practical questions to every PhD student. Although some experienced students learned expertises from previous courses or practices, an empirical pattern is fairly appreciated. Secondly, the state-of-art tools are not efficient enough for collaborative work in this use case. Since the qualifying paper itself evolves and changes lifecycle in a distributed production environment, several versions generate, and various comments and reviews mix. A supervisor could give some general comments by email, while commenters and reviewers suggest several detailed critiques or referenced materials with un-unified formats of files. There is still no standard schema and container to describe, comment on, and review scientific papers to facilitate collaboration, version management and metadata sharing. Thirdly, when the student hunts for background knowledge about his research topic, it frequently happens that he desires to gain some most interesting parts of references for further in-depth reading directly, such as a result of an evaluation experiment, a definition of a novel concept, and an impressed figure, etc. To date, a scientific publication is always applied as a basic indivisible unit such as a pdf document, which needs a specific modularity for a paper's rhetorical structure and interlinked knowledge representation. Fourthly, when the student finds some interesting related works, e.g. a reference, a relevant project, or even a researcher mentioned in a paper, he has to input their titles or names to the search engines beginning with a time-consuming navigation. Even thus, months' work effort can just result a 10-page paper without his elaborate knowledge collections, which will dramatically benefit others in case of sharing. Instead of perishing, marking them up as entities and annotating them with Uniform Resource Identifiers (URI), along

[3] Google Scholar: `http://scholar.google.com/`

[4] Citeseer: `http://citeseer.ist.psu.edu/`

[5] DBLP: `http://dblp.uni-trier.de/`

with sets of attributes could definitely facilitate the efficiency for article search and navigation. Semantically enriching papers is still a difficult problem, yet to be adequately resolved. Papers always lack semantics both during authoring and post-publication period. To help readers easily and intuitively attain a rhetorical block which describes background, contribution or discussion is another research issue to be tackled.

The methodology we are applying is dedicated to enriching semantics and empirical knowledge into the whole lifecycle of scientific publications. Preliminary solution proves possible feasibilities of semantically managing ubiquitous Scientific Knowledge Objects (SKO) during their creation, evolution, collaboration and dissemination. Also, we aim to provide a viable means to generate semantic documents for scientific publications in a simple and intuitive way [1].

To achieve this objective, we have attempted to introduce a SKO Patterns framework [2] including SKO Metadata Schema, SKO Patterns Repository and SKO Editor three components, which are dedicated to resolve the problems addressed above. SKO Metadata [3] is an extension of current standards e.g. Dublin Core[6] and Learning Object Metadata (LOM) [4] in digital library area, which supplies the schema with semantics and lifecycle features. SKO Patterns Repository uses the methodology of pattern language and is based on three ontologies, i.e. document ontology, annotation ontology and rhetorical ontology. It's a faceted classification, dealing with not only syntactic patterns, but also semantic ones. In current stage of our project, we mainly focus on the rhetorical ontology and the other two will be implemented in the future work. SKO Editor supports a semantic editing environment for managing SKOs and their metadata during both authoring and post-publication [5]. These three components constitute the foundation for SKO Patterns theory and applications.

Acknowledgement

This research is partially supported by China Scholarship Council (CSC). And many thanks to Prof. Fausto Giunchiglia for his inspiring suggestions.

References

[1] Xu, H., Zhang, C.: Towards a pattern-based approach for scientific writing and publishing in chinese. In: Chowdhury, G., Koo, C., Hunter, J. (eds.) The Role of Digital Libraries in a Time of Global Change. LNCS, vol. 6102, pp. 264–265. Springer, Heidelberg (2010)

[2] Xu, H.: A semantic pattern approach to managing scientific publications. In: Kim, T.-h., Adeli, H. (eds.) AST/UCMA/ISA/ACN. LNCS, vol. 6059, pp. 431–434. Springer, Heidelberg (2010)

[6] Dublin Core: `http://dublincore.org/`

[3] Xu, H.: Managing ubiquitous scientific knowledge on semantic web. In: Kim, T.-h., Adeli, H. (eds.) AST/UCMA/ISA/ACN. LNCS, vol. 6059, pp. 421–430. Springer, Heidelberg (2010)

[4] IEEE learning technology standards committee (ltsc). IEEE P1484.12 Learning Object Metadata Working Group (2000)

[5] Xu, H.: A pattern-based representation approach for online discourses. In: Kim, T.-h., Adeli, H. (eds.) AST/UCMA/ISA/ACN. LNCS, vol. 6059, pp. 378–384. Springer, Heidelberg (2010)

A General Bayesian Network-Assisted Ensemble System for Context Prediction: An Emphasis on Location Prediction

Kun Chang Lee[1] and Heeryon Cho[2]

[1] Professor of MIS and WCU Professor of Creativity Science,
SKK Business School and Department of Interaction Science,
Sungkyunkwan University, Seoul 110-745, Republic of Korea
kunchanglee@gmail.com, leekc@skku.edu
[2] Department of Interaction Science,
Sungkyunkwan University, Seoul 110-745, Republic of Korea
heeryon@gmail.com, heeryon@skku.edu

Abstract. Context prediction, highlighted by accurate location prediction, has been at the heart of ubiquitous decision support systems. To improve the prediction accuracy of such systems, various methods have been proposed and tested; these include Bayesian networks, decision classifiers, and SVMs. Still, greater accuracy may be achieved when individual classifiers are integrated into an ensemble system. Meanwhile, General Bayesian Network (GBN) classifier possesses a great potential as an accurate decision support engine for context prediction. To leverage the power of both the GBN and the ensemble system, we propose a GBN-assisted ensemble system for location prediction. The proposed ensemble system uses variables extracted from Markov blanket of the GBN's class node to integrate GBN, decision tree, and SVM. The proposed system was applied to a real-world location prediction dataset, and promising results were obtained. Practical implications are discussed.

Keywords: Context Prediction, Location Prediction, Ensemble Methods, General Bayesian Network, GBN-Assisted Ensemble Classifier, ID3, C4.5, CART, SVM.

1 Introduction

Ubiquitous decision support systems [1] adapt to users' changing contexts to provide context-sensitive services that meet users' needs and preferences. Such adaptation requires the system to predict user's future context based on the current context data gathered from various sources, e.g., user's handheld devices, sensors embedded in the environment, and distributed databases. Context is defined as, "any information that can be used to characterize the situation of an entity. An entity is a person, place or object that is considered relevant to the integration between a user and an application, including the user and the application themselves [2]." The collected (context) data are used as training and test data for building context prediction models, and the constructed models are used to infer user's future context. The success of context prediction depends on the degree of accuracy of related predictors.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 294–303, 2010.

Existing works have successfully employed Bayesian networks to predict future contexts: Patterson et al. [3] used dynamic Bayesian network to predict likely travel destinations on a city map; Hwang and Cho [4] proposed a modular Bayesian network model to infer landmarks of users from mobile log data collected through smart phones; and Kasteren and Kröse [5] used naive Bayesian and dynamic Bayesian networks to infer daily activities of elderly people performed inside their houses; Sánchez et al. used discrete hidden Markov model to automatically estimate hospital-staffs' activities [6].

Other works have employed decision tree classifiers and Support Vector Machines (SVM) for context prediction: Byun and Cheverst [7] used decision tree to infer the preferences of the user in an intelligent office environment; Lum and Lau [8] proposed a negotiation algorithm based on decision trees to handle content adaptation for mobile devices; and Matsuo et al. [9] used SVM to infer user's long-term properties such as gender, age, profession, and interests from location information; their system automatically learns patterns between users' locations and user properties.

While individual prediction models perform well on many context prediction tasks, better performance may be achieved by harnessing the power of multiple models. Ensemble-based systems, also known as multiple classifier systems, committee of classifiers, or mixture of experts, combine individual classifiers that make errors on different parts of data to enhance prediction performance. This paper investigates yet another kind of context prediction using General Bayesian Network (GBN)-based ensemble system; we investigate the power of GBN-*assisted* ensemble system on location prediction.

A GBN-assisted ensemble system is defined as a multiple classifier system that uses those variables inside the Markov blanket of a GBN's class node (or target node) to build an ensemble system. A Markov blanket of a node X consists of the direct parent of X, the direct successors of X, and all direct parents of X's direct successors [10] in a given Bayesian network. The Markov blanket of node X may be thought of as the minimal set of nodes that isolates X from the rest of the graph [11]. If a node is absent from the target node's Markov blanket, its value is completely irrelevant to the prediction [12]. Hence, the Markov blanket can be used to select the core variables that affect the class variable. In this paper, we first create a GBN to identify the variables inside the Markov blanket of GBN's class node, and then use those selected variables to create GBN-assisted ensemble system by merging GBN, decision tree, and/or SVM using voting, stacking, and grading combination strategies. Location prediction experiments are conducted using real-world data to evaluate the prediction accuracies of GBN-assisted two-classifier and three-classifier ensemble systems.

The contributions of this paper are twofold: (1) the finding that the GBN-assisted ensemble systems are comparable to the GBN-based ensemble systems; (2) what practical implications the GBN-assisted ensemble system has for location prediction. Section 2 describes three types of classifier combination strategies used for building ensemble classifiers for the location prediction experiment. Section 3 describes data, the experimental setup, and the experimental results of the ensemble performance evaluation. Practical implications of the GBN-assisted ensemble approach are discussed in Section 4. Conclusion and future research issues are given in Section 5.

2 Ensemble Methods: Voting, Stacking, and Grading

Many studies have shown that fusing a set of different classifiers (or ensemble of classifiers) with different misclassified instances (i.e., ones that do not overlap) will yield better classification performance over an individual classifier, which makes up the ensemble system, having the best performance [13]. The intuition is that if different classifiers make errors on different instances, the strategic combination of these classifiers can reduce the overall error to improve the performance of the ensemble system [14].

The success of ensemble system depends on achieving diversity among individual classifiers with respect to misclassified instances. There are four ways to achieve this diversity [14]: (1) use different training examples to train individual classifiers; (2) use different training parameters; (3) use different features to train classifier; or (4) combine entirely different type of classifiers. The first approach deals with incorporating various resampling techniques; bagging (or bootstrap aggregating) [15] and boosting [16] are two well known techniques. The second approach deals with using different parameter values such as weights, nodes, or layers (depending on the classifier to be trained) to train the individual classifier. The third approach deals with using different features to train the classifier; random subspace method [17] is one such method. Finally, the last approach deals with combining entirely different type of classifiers; an example would be combining decision trees, SVMs, and nearest neighbor classifiers.

In this paper, we focus on the last approach of combining entirely different type of classifiers to construct ensemble systems for location prediction experiment. We select three types of individual classifiers – decision trees, Bayesian classifiers, and SVM – and integrate them using three different combination strategies – voting, stacking, and grading. Note that prior to ensemble system creation, we first create GBN to identify the variables inside the Markov blanket of the class node; these variables, which parsimoniously describe the class node, are used to create GBN-assisted ensemble systems. We now briefly introduce each combination strategy.

Voting. Voting or majority voting has been used for centuries by humans to make decisions. The same methodology is employed to determine the final outcome of multiple classifier system. Three versions of majority voting exist [18]: unanimous voting in which all agree to the final decision, simple majority voting in which the final decision exceeds 50% + 1 votes, and plurality voting in which one with the most votes becomes the final decision. While these approaches cast entire vote to a single class that each classifier considers most likely, voting can also combine classifiers by averaging each classifier's probability estimates. We average each classifier's probability estimates when using voting strategy in our experiment.

Stacking. Stacking or stacked generalization [19] uses a high-level method (called level-1 generalizer or meta-learner) to combine lower-level methods (called level-0 models or base classifiers). The predictions of the lower-level methods are used as training data for high-level method. The ensemble learning proceeds in two steps: first, the predictions of level-0 models are calculated; then, the predictions are used as training data for training level-1 generalizer. The class labels of the original data are retained for level-1 learner's training data. In essence, stacking provides the meta-learner indirect feedback about the correctness of its base classifiers [20].

Grading. Grading [20] uses "graded" predictions (i.e., whether the prediction is correct, marked as "+", or incorrect, marked as "-") of the base classifiers to train a meta-classification scheme. The learning proceeds in two-steps first by obtaining correct/incorrect predictions of the base classifiers, and then replacing original class values with correct/incorrect predictions to learn meta-classification schemes. When a new instance is tested, each base classifier makes a prediction, and the final prediction is derived from the base classifiers that are predicted to be correct by the meta-classification schemes. If conflicts exist within several base-level predictions, they can be resolved by voting and other methods.

Existing researches have mainly applied ensemble-based system to credit scoring analysis [21-24], bankruptcy prediction [25], heart disease diagnosis [26, 27], and traffic incident detection [28]. Some have focused on location prediction [29]. We also apply our GBN-assisted ensemble approach to the problem of location prediction.

Table 1. Variables, variable values, and number of values in the location prediction dataset

Variable	Variable Value (No. of Values)
Location Departed	600thAnniversaryBuilding, BasketballCourt, Bicheondang, BusinessBuilding, CentralLibrary, DasanHallOfEconomics, EastGate, FacultyHall, FrontGate, GeumjandiSquare, HoamHall, InternationalHall, LargePlayground, LawBuilding, Myeongnyundang, Oacknyujeong, OutsideCampus, RearGate, StudentUnion, SuseonHall, SuseonHallAnnex, ToegyeHallOfHumanities, Yanghyeongwan, Yurimhoegwan (24)
Path Start	A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, Y (24)
Path Middle	A, AB, ABG, B, BA, BC, BE, BF, BFJ, BG, BGI, BGJ, BGM, BH, BHI, BQ, BQJ, C, CB, CBG, D, E, EBG, F, FB, G, GB, GBA, GJ, GM, H, HB, I, IJ, IM, J, JF, JFB, JG, JGB, JI, JK, JQB, K, M, MGB, MJ, N, NJ, none, Q, QB, QBA, SGB, T, X, XM (57)
Path End	A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y (25)
Location Arrived	600thAnniversaryBuilding, BasketballCourt, Bicheondang, BusinessBuilding, CentralLibrary, DasanHallOfEconomics, EastGate, FacultyHall, FrontGate, GeumjandiSquare, HoamHall, InternationalHall, InternationalHouse, LargePlayground, LawBuilding, Oacknyujeong, OutsideCampus, RearGate, StudentUnion, SuseonHall, SuseonHallAnnex, ToegyeHallOfHumanities, Yanghyeongwan, Yurimhoegwan (24)
Activity	ChatWithFriends, ClubActivity, Consult, Eat, Exercise, FinancialErrands, Hobby, Homework, InternetSearch, JobHunting, Lecture, MiscErrands, Other, Part-TimeJob, Shop, Study, TeaTime (17)
Major	BizAdmin, Confucianism, DomesticScience, Economics, Education, Engineering, FineArts, FreeMajor, InfoTechnology, Law, LiberalArts, SocialScience, SportsScience (13)
Student Year	Freshman, Sophomore, Junior, Senior (4)
Gender	Male, Female (2)
Weekday Leisure	Concert/Exhibitions, Games, IndividualSports, Socialize, TeamSports, Travel (6)
Lunch Time Leisure	
Monthly Allowance	<100, 100-300, 300-500, 500-700, 700-900, >900 (6)

3 Empirical Evaluation

To investigate the location prediction performance of the GBN-assisted ensemble system, we collected real-world data from undergraduate students, constructed two-classifier and three-classifier ensemble systems, and evaluated their performances. Hereafter, we indicate ensemble systems as ensembles or ensemble classifiers.

3.1 Data

User context data were collected from 335 college students in Seoul, Korea to create training and test data for location prediction experiment. There were 205 male students and 130 female students, and the student year composition was as follows: 131 freshmen, 38 sophomores, 64 juniors, and 102 seniors. The students were asked to complete a demographic survey which asked his/her gender, major, student year, weekday leisure activity, lunch-time leisure activity, monthly allowance, and student ID. Then, they were instructed to document their whole-day activity on campus for any two days; in particular, where they visited via what route and what activity they engaged in at the visited location. A list of campus location codes, route codes (a list of letters was specified to describe a sequence of paths), and activity codes were provided to help them record their activities. They were given extra credits for their work.

After the two types of data (i.e., the demographic data and the campus activity data) were cleaned, they were merged using the student ID to create a campus activity-demographic data. The merged data contained 12 variables: ‘Location Arrived’, ‘Path Start’, ‘Path Middle’, ‘Path End’, ‘Location Departed’, ‘Activity’, ‘Gender’, ‘Major’, ‘Year’, ‘Weekday Leisure’, ‘Lunch Time Leisure’, and ‘Monthly Allowance’. Table 1 lists the 12 variables and the values of each variable. A total of 3,150 records of campus activity-demographic data were used in the experiment. The ‘Location Arrived’ variable was selected as the class variable in the experiment.

3.2 Experimental Setup

WEKA [30], an open source data-mining tool, was used to construct and evaluate the GBN-based ensembles and the GBN-assisted ensembles. The GBN-based ensembles were built using all 12 variables of the campus activity-demographic data whereas the GBN-assisted ensembles used only 5 (‘Location Arrived’, ‘Path End’, ‘Location Departed’, ‘Activity’, and ‘Major’); these 5 variables were selected on the basis of the Markov blanket of GBN’s class node. That is, prior to creating the GBN-assisted ensembles, a GBN was created using the 12-variable data to identify the variables that parsimoniously describe the class variable. Note that the class variable is included in the 12-variable and 5-variable dataset.

Three types of decision trees (ID3 [31], C4.5 [32] or J48 in WEKA, and CART [33]), two types of Bayesian network classifiers (GBN-K2 [34] and GBN-Hill Climb, hereafter GBN-HC), and one SVM [35] were integrated to create two-classifier (GBN+DecisionTree or GBN+SVM) and three-classifier (GBN+DecisionTree+SVM) ensemble classifiers. Each ensemble classifier employed voting, stacking, and grading strategy. All in all, 24 two-classifier (Table 3) and 18 three-classifier (Table 4) ensembles were created for each of the GBN-based and the GBN-assisted approaches.

All algorithms needed for creating the individual classifiers and the ensemble classifiers were already available in WEKA version 3.6.2. For the GBN-K2 and the GBN-HC classifier construction, the maximum number of parent node was set to 2 and the BAYES scoring metric was used. For the decision trees (ID3, J48, CART) and the SVM (SMO algorithm), the default settings in WEKA were used. As for the three combination strategies, the averaging of the probability estimates was used to combine classifiers for voting. For level-1 generalizer (or meta-classifier) for stacking, instance-based learning algorithm (IB*k* [36]) was used with ten nearest neighbors following [20]. The same meta-classification scheme was used for grading.

We performed one run of 10-fold cross-validation on each ensemble classifier to obtain its prediction accuracy, and then conducted paired t-tests at the 1% and 5% significance level to compare each ensemble classifier with the baseline GBN-K2 individual classifier.

3.3 Results

Table 2 compares the prediction accuracies of individual classifiers created using the 12 variables (first row) and 5 variables (second row). The SVM shows the best individual-classifier prediction accuracy in both cases; ID3 shows the worst. Compared to the GBN-K2 individual classifier (84.38), only the SVM created using 5 variables shows significantly better performance at the 5% significance level.

Table 2. Prediction accuracies of 12-variable and 5-variable individual classifiers compared to GBN-K2 individual classifier. (* p<0.05)

Classifier	GBN-K2	GBN-HC	ID3	J48	CART	SVM
12-variable	84.38	81.46	80.67	84.73	83.97	85.14
5-variable	84.38	84.79	79.14	83.08	83.71	**85.52***

Table 3 compares the prediction accuracies of the GBN-based (12 variables are used) two-classifier ensembles versus the GBN-assisted (5 variables inside the Markov blanket of the GBN's class node are selected and used) two-classifier ensembles. The prediction accuracies showing statistically better performance to the GBN-K2 individual classifier are marked in bold in the tables. Compared to the GBN-K2 individual classifier, the GBN-K2+ID3 voted-ensemble (85.56, 5-variable) and the GBN-K2+J48 graded-ensemble (85.90, 12-variable) show significantly better performance at the 1% significance level; six and three additional ensemble classifiers show better performance at the 5% significance level for the GBN-based (12-variable) and the GBN-assisted (5-variable) approaches, respectively. Overall, the GBN+ID3 voted-ensembles and GBN+SVM graded-ensembles show good performances. It is notable that GBNs can benefit from the lowest-performing ID3 individual classifier; conversely, some classifier combination can hurt the performance as shown in the case of the GBN-K2+SVM stacked-ensemble (84.22 and 83.78).

Table 3. Prediction accuracies of GBN-based (12-variable) and GBN-assisted (5-variable) 2-classifier ensembles compared to GBN-K2 individual classifier. (* p<0.05, ** p<0.01)

2-Classifier Ensemble		Voting		Stacking		Grading		Average	
		12-var	5-var	12-var	5-var	12-var	5-var	12-var	5-var
GBN-K2	ID3	**85.97***	**85.56****	85.17	84.98	84.67	84.41	85.27	84.98
	J48	**85.68***	84.48	**85.84***	84.51	**85.90****	84.98	85.81	84.66
	Cart	84.63	84.41	84.00	83.75	84.79	84.57	84.47	84.24
	SVM	84.48	84.44	84.22	83.78	**85.62***	**85.11***	84.77	84.44
GBN-HC	ID3	**85.97***	**85.62***	84.51	84.95	82.92	84.35	84.47	84.97
	J48	84.60	84.10	84.03	84.95	84.22	84.98	84.28	84.68
	Cart	84.22	84.73	84.06	84.16	84.10	84.70	84.13	84.53
	SVM	81.56	84.79	82.48	84.35	**85.43***	**85.56***	83.16	84.90
Average		84.64	84.77	84.29	84.43	84.71	84.83	84.54	84.68

Table 4. Prediction accuracies of GBN-based (12-variable) and GBN-assisted (5-variable) 3-classifier ensembles compared to GBN-K2 individual classifier. (* p<0.05, ** p<0.01)

3-Classifier Ensemble		Voting		Stacking		Grading		Average	
		12-var	5-var	12-var	5-var	12-var	5-var	12-var	5-var
GBN-K2 + SVM	ID3	**86.00***	**85.52****	85.30	84.63	**86.16****	85.17	85.82	85.11
	J48	**85.71***	84.48	85.56	84.70	**85.90****	**85.71****	85.72	84.96
	Cart	84.67	84.41	84.10	83.68	**85.87***	**85.30***	84.88	84.46
GBN-HC + SVM	ID3	**85.97***	**85.59***	85.05	84.98	**85.87***	85.24	85.63	85.27
	J48	84.60	84.16	83.94	84.51	**85.87****	**85.62****	84.80	84.76
	Cart	84.22	84.73	83.56	83.52	85.49	85.08	84.42	84.44
Average		85.20	84.82	84.59	84.34	85.86	85.35	85.21	84.84

Table 4 shows the prediction accuracies of the GBN-based (12-variable) three-classifier ensembles versus the GBN-assisted (5-variable) three-classifier ensembles. The prediction accuracy increases in general for the graded ensemble approach when an SVM classifier is added to the GBN+DecisionTree ensembles. Compared to the GBN-K2 individual classifier, three GBN-based and three GBN-assisted ensembles show significantly better prediction performance at the 1% significance level and five GBN-based and two GBN-assisted ensembles show better performance at the 5% significance level. The voting strategy seems to work well with the ID3-included ensembles and the grading strategy with the J48-included ensembles regardless of the GBN-based or GBN-assisted approaches. On the contrary, stacking, as a combination strategy, does not seem to work well; of the twelve GBN+DecisionTree+SVM stacked-ensembles, seven show worse performance than their GBN+DecisionTree two-classifier counterparts. For instance, the GBN-HC+J48+SVM and GBN-HC+CART+SVM ensembles both display lower prediction performance than their GBN+DecisionTree two-classifier counterparts regardless of the GBN-based or GBN-assisted approaches. In general, the grading strategy turned out to be the best ensemble combination strategy for the location prediction dataset in this paper, the voting strategy turned out to be the second best, and the stacking strategy turned out to be the worst strategy.

4 Discussion

In the experiment, we were able to confirm that the prediction accuracies of 24 ensemble classifiers (Tables 3 and 4, numbers in bold) were statistically better than the baseline GBN-K2 individual classifier. But how do they compare to one another? To see whether any of these superior ensemble classifiers perform significantly better than the other, we conducted a one-way ANOVA using a significance level of $\alpha = 0.05$ to check the statistical differences in prediction accuracy. The result showed that the differences among the prediction accuracies of 24 ensemble classifiers were statistically insignificant ($F(23, 216) = .209$, $p = 1.000$); no one ensemble classifier outperformed the other. With no clear winner present, choosing a two-classifier ensemble (Table 3, any one in bold) over a three-classifier ensemble (Table 4, any one in bold) could be a rational decision since the two-classifier ensembles require lower computational cost while maintaining comparable prediction performance to the three-classifier ensembles. Similar rationale could be applied to the GBN-based (12-variable) approach versus the GBN-assisted (5-variable) approach; with less number of features (variables) to work on, choosing the GBN-*assisted* approach could save on computational cost without hurting the prediction performance; but this is not the only advantage of the GBN-assisted approach.

Because the GBN-assisted ensemble approach exploits the Markov blanket of the target node, context prediction can be performed efficiently by focusing on the truly relevant explanatory variable(s); the GBN-assisted ensemble approach can be said to encapsulate a natural feature selection capability that identifies the features that parsimoniously describe the target variable. Such feature selection (or reduction) capability is beneficial to both humans and machines. For example, when designing a context prediction system, data gathering strategy for context prediction must be coordinated. If the prediction model inside the system requires too many context data to predict future context, both the users and the system will need to make much effort in handling these data. Keeping the number of features to a minimum lowers the data-handling cost for both the users and machines, and it also keeps the model simple.

Although the ensemble systems sacrifice model interpretability over performance, both the GBN-based ensemble approach and the GBN-assisted ensemble approach allow humans to understand the variable relationship through individual GBN models. A GBN expresses the relationship between a target variable and explanatory variables using nodes and links; humans can easily interpret how variables influence each other through this graph model. Since humans can understand which explanatory variables directly influence the target variable in the GBN, the graph model can be used in what-if and goal-seeking analyses. A what-if analysis is one in which decision makers analyze the possible results by considering intended changes to input conditions. A goal-seeking analysis is closely related to such simulation activities in which a certain goal is suggested, and decision makers attempt to observe what kind of input conditions are necessary to obtain such a goal. Such capabilities are the advantages of the GBN-related ensemble approaches. Lastly, it should be noted that although the stacking approach discussed in this paper showed a poor performance, changing the meta-level learner and parameter setting may improve the performance.

5 Concluding Remarks

A GBN-assisted ensemble system exploits the Markov blanket of the GBN's target node to identify and select the core features. We compared the prediction performance of the GBN-assisted approach, which uses fewer variables, to the GBN-based approach, which uses more variables, and found that the performance of the two approaches were comparable to each other despite the fact that the GBN-assisted approach handled fewer features. In this sense, we can view the GBN-assisted ensemble approach to have a computational edge over the GBN-based ensemble approach. Ensemble systems generally have better prediction accuracy, but have low interpretability of the models. The GBN-related ensemble approaches can compensate for the sacrificed interpretability by exploiting and exploring the variable relationship depicted in the individual GBN graph model. In the future, we plan to design a context prediction system which can handle multiple prediction models suited to various user groups; we plan to place the GBN-assisted ensemble systems at the heart of the system.

Acknowledgments. This research was supported by the Ubiquitous Computing and Network (UCN) Project, Knowledge and Economy Frontier R&D Program of the Ministry of Knowledge Economy (MKE) in Korea as a result of UCN's subproject 09C1-T2-20S.

References

1. Kwon, O., Yoo, K., Suh, E.: UbiDSS: A Proactive Intelligent Decision Support System as an Expert System Deploying Ubiquitous Computing Technologies. Expert Systems with Applications 28, 149–161 (2005)
2. Dey, A.K.: Understanding and Using Context. Personal and Ubiquitous Computing 5, 4–7 (2001)
3. Patterson, D., Liao, L., Fox, D., Kautz, H.: Inferring High-Level Behavior from Low-Level Sensors. In: Dey, A.K., Schmidt, A., McCarthy, J.F. (eds.) UbiComp 2003. LNCS, vol. 2864, pp. 73–89. Springer, Heidelberg (2003)
4. Hwang, K.S., Cho, S.B.: Landmark Detection from Mobile Life Log Using a Modular Bayesian Network Model. Expert Systems with Applications 36, 12065–12076 (2009)
5. van Kasteren, T., Kröse, B.: Bayesian Activity Recognition in Residence for Elders. Intelligent Environments, 209–212 (2007)
6. Sánchez, D., Tentori, M., Favela, J.: Activity Recognition for the Smart Hospital. IEEE Intelligent Systems 23, 50–57 (2008)
7. Byun, H.E., Cheverst, K.: Utilizing Context History to Provide Dynamic Adaptations. Applied Artificial Intelligence 18, 533–548 (2004)
8. Lum, W.Y., Lau, F.C.M.: A Context-Aware Decision Engine for Content Adaptation. IEEE Pervasive Computing 1, 41–49 (2002)
9. Matsuo, Y., Okazaki, N., Izumi, K., Nakamura, Y., Nishimura, T., Hasida, K., Nakashima, H.: Inferring Long-Term User Properties Based on Users' Location History. In: 20th International Joint Conference on Artificial Intelligence (IJCAI), pp. 2159–2165 (2007)
10. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo (1988)
11. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)

12. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco (2005)
13. Kittler, J.: Combining Classifiers: A Theoretical Framework. Pattern Analysis & Applications 1, 18–27 (1998)
14. Polikar, R.: Ensemble Based Systems in Decision Making. IEEE Circuits and Systems Magazine 6, 21–45 (2006)
15. Breiman, L.: Bagging Predictors. Machine Learning 24, 123–140 (1996)
16. Schapire, R.E.: The Strength of Weak Learnability. Machine Learning 5, 197–227 (1990)
17. Ho, T.K.: The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 832–844 (1998)
18. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience, New Jersey (2004)
19. Wolpert, D.H.: Stacked Generalization. Neural Networks 5, 241–259 (1992)
20. Seewald, A., Fürnkranz, J.: An Evaluation of Grading Classifiers. In: Advances in Intelligent Data Analysis, pp. 115–124 (2001)
21. Hsieh, N.C., Hung, L.P.: A Data Driven Ensemble Classifier for Credit Scoring Analysis. Expert Systems with Applications 37, 534–545 (2010)
22. Zhou, L., Lai, K.K., Yu, L.: Least Squares Support Vector Machines Ensemble Models for Credit Scoring. Expert Systems with Applications 37, 127–133 (2010)
23. Twala, B.: Multiple Classifier Application to Credit Risk Assessment. Expert Systems with Applications 37, 3326–3336 (2010)
24. Yu, L., Yue, W., Wang, S., Lai, K.K.: Support Vector Machine Based Multiagent Ensemble Learning for Credit Risk Evaluation. Expert Systems with Applications 37, 1351–1360 (2010)
25. Hung, C., Chen, J.H.: A Selective Ensemble Based on Expected Probabilities for Bankruptcy Prediction. Expert Systems with Applications 36, 5297–5303 (2009)
26. Das, R., Sengur, A.: Evaluation of Ensemble Methods for Diagnosing of Valvular Heart Disease. Expert Systems with Applications 37, 5110–5115 (2010)
27. Eom, J.H., Kim, S.C., Zhang, B.T.: AptaCDSS-E: A Classifier Ensemble-Based Clinical Decision Support System for Cardiovascular Disease Level Prediction. Expert Systems with Applications 34, 2465–2479 (2008)
28. Chen, S., Wang, W., Van Zuylen, H.: Construct Support Vector Machine Ensemble to Detect Traffic Incident. Expert Systems with Applications 36, 10976–10986 (2009)
29. Anagnostopoulos, T., Anagnostopoulos, C., Hadjiefthymiades, S., Kyriakakos, M., Kalousis, A.: Predicting the Location of Mobile Users: A Machine Learning Approach. In: Int'l Conf. Pervasive Services, pp. 65–72 (2009)
30. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. ACM SIGKDD Explorations Newsletter 11, 10–18 (2009)
31. Quinlan, J.R.: Induction of Decision Trees. Machine Learning 1, 81–106 (1986)
32. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993)
33. Breiman, L.: Classification and Regression Trees. Chapman & Hall/CRC (1984)
34. Cooper, G.F., Herskovits, E.: A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine Learning 9, 309–347 (1992)
35. Platt, J.C.: Fast Training of Support Vector Machines Using Sequential Minimal Optimization. In: Advances in Kernel Methods: Support Vector Learning, pp. 185–208. MIT Press, Cambridge (1999)
36. Aha, D.W., Kibler, D., Albert, M.K.: Instance-Based Learning Algorithms. Machine Learning 6, 37–66 (1991)

Bayesian Network Approach to Predict Mobile Churn Motivations: Emphasis on General Bayesian Network, Markov Blanket, and What-If Simulation

Kun Chang Lee[1] and Nam Yong Jo[2]

[1] Professor of MIS and WCU Professor of Creativity Science
SKK Business School and Department of Interaction Science
Sungkyunkwan University
Seoul 110-745, Republic of Korea
kunchanglee@gmail.com, leekc@skku.edu
[2] PhD Candidate
SKK Business School
Sungkyunkwan University
Seoul 110-745, Republic of Korea
namyong.jo@gmail.com

Abstract. As mobile telecommunication service becomes indispensable to our daily life, predicting the reasons of mobile churn seems essential from the perspective of mobile service providers. Previous studies have been focused on mobile churn prediction itself, not churn motivations which can play as a good indicator to forecasting real churn. Therefore, main focus of this study is placed on predicting mobile churn motivations, instead of mobile churn prediction. We propose BN approach to predict mobile churn motivation, adopting three types of BN models such as Naïve BN (NBN), Tree Augmented NBN (TAN), and General BN (GBN). To prove its validity in predicting mobile churn motivations, benchmarking classifiers were adopted and their performance was compared with BN classifiers. Through analyzing the empirical results, we found three advantages of GBN-(1) GBN performance is competitive compared with other benchmarking classifiers, (2) Markov Blanket (MB) variables are considerably small in number and make it handy for decision makers, and (3) what-if simulation is possible, which is not possible in other benchmarking classifiers. Practical implications of empirical results were addressed.

Keywords: Mobile Churn Prediction, Churn Motivation, Bayesian Network, Markov Blanket, What-If Simulation.

1 Introduction

The mobile telecommunication industry is a cut-throat world. Mobile customers seek out emerging devices and services. The annual churn rate ranges from 20% to 40% for most global mobile telecommunications service companies [9]. Reducing churn is important because acquiring new customers is more expensive than retaining existing

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 304–313, 2010.

customers [17]. In order to manage customer churn to increase profitability, companies need to predict churn behavior, however this problem not yet well understood [1].

Thus, to address churn, we propose a new approach based on Bayesian network (BN). First, we present a BN structured learning algorithm to uncover the underlying motivations of churn. Second, we propose the concept of a Markov Blanket (MB) as a robust feature (variable) selection method to make our model more parsimonious. Forth, benchmarking test is performed to compare BNs with other popular classifiers. Finally, we identify the underlying churn motivations and integrate the decision-making procedure. Insight on motivations of customers churn is gained by interpreting the probabilities in these causal prediction models.

The paper is organized as follows: Section 2 contains a brief overview of related works and introduces the BN prediction models. Section 3 reviews the overall framework, a detailed description of the experiments and the results. Finally, the major contributions of the paper, managerial implications, and future research directions are discussed in Section 4.

2 Theoretical Background

2.1 Mobile Churn

Churn occurs when a customer terminates the use of a service from the service provider either voluntarily or involuntarily. In the telecommunication market, churn can be measured as the cancelation rate in a certain period of time [17]. Studies on predicting churn rate in the telecommunication market often have used Logit and Probit statistics methods. Recently, many artificial intelligence methods such as neural networking (NN) and decision trees (DT) have been used to study churn. Contractual variables and phone call-related variables were used to evaluate the explanatory variables of churn [18].

Most researches in mobile domain can be categorized into two kinds of styles. The first category focuses on causal variables to determine customer satisfaction or usage pattern through logistic regression (Logit) or structural equation models (SEM) using survey data [1] [3]. The second category relates to churn prediction, especially to determine better classifiers which have better classification accuracy [11] [13]. Thus, prior research reveals that the primary focus of previous churn prediction models has been limited to maximizing predictability, with little attention given to the issues of the motivation of customers' churn.

2.2 Bayesian Network

A Bayesian network (BN) can be defined as a directed acyclic graph (DAG) which has a probabilistic causal relationship and direction. Bayesian networking has been used effectively as a tool to support decision making in finance and marketing. [16] used BN to develop an early warning system for bankruptcy, and [7] utilized BN to increase the quality of loan assessment.

From the structural point of view, the most general type of BNs is Naïve Bayesian Network (NBN), Tree Augmented Naïve Bayes (TAN) and General Bayesian network (GBN). NBN shows a shape in which a class node is linked with all of the children nodes which are explanatory variables for the target variable. All of the explanatory attributes are dependent on the class node but independent on each other. In NBN, the class node is a special variable distinguished from other nodes, and the model assumes too much independence between variables. Therefore, it has been regarded as too rigid and not a proper reflection of reality, leading to the introduction of TAN by Friedman et al. [6] to compensate for those weaknesses. TAN expands NBN into a tree shape. Unlike the first two BNs, the GBN does not allow any differences between class nodes and general attribute nodes to express inter-dependency in the Bayesian network [2]. Therefore, class nodes can also have parent nodes, and the causal relationships or inter-dependencies among various variables in a given decision-making issue can be most naturally expressed. Rather than searching for a fully unrestricted general BN structure, as in an unsupervised Bayesian learning algorithm, the GBN classifier can seek to optimize classification performance based primarily on the probabilistically salient Markov blanket nodes.

In this paper, we use all three types of BN structure to compare performance as well as to reveal the underlying structure of churn motivations. In addition, only variables in the Markov blanket are used for further analysis to see if performance of GBNs is still competing. The concept of the Markov blanket (MB) was first introduced by Pearl [14] but has recently received renewed attention in the areas of Bayesian learning [4] and features selection [10]. The probabilistic nature of the Markov blanket can be explained using the concept of *d-separation* (*direction dependent separation*) [14], a graphical criterion related to the blocking of information flow among variables. In a faithful Bayesian network, *d-separation* captures all of the conditional independence relationships encoded in the network [14]. If all variables in MB of a node are instantiated, then the node is *d-separated* from the rest of the network. In other words, *MB*(*T*) is a minimal feature subset required to predict *T*, which graphically corresponds to a set neighborhood of *T*: its parents, its children and the other parents of its children.

3 Experiments

3.1 Data and Variables

The data in this paper were donated by a major mobile telecommunication company in South Korea. This is originally consisted of 14 variables and 5,000 records that were randomly sampled from anonymous churned customers. After removing records containing a missing value 4,922 records are remained for experiments in this research.

Table 1 summarizes the variables used in this study. Customers' age (Age), makers of the mobile devices (DeviceMaker), customers' service grades provided by companies (ServiceGrade) and the way customers pay the bill (Paymentway) are variables originally donated by categorized. Continuous variables reflect the number of calls to the contact center in a previous month (CallsInaMon), average revenue per user in a previous month (ARPUInaMon), average revenue per user over previous three months (ARPUIn3MonAve), the frequencies of voice calls in a previous month (VoiceFreqInaMon) and the minutes of voice usage in a previous month (VoiceMinuInamon). These variables were meaningfully transformed into categorical variables using a supervised discrete method supported by WEKA[1]. We inserted two more additional

[1] Available at `http://www.cs.waikato.ac.nz/ml/weka/`

continuous variables to characterize customers' loyalty such as months of usage (totMonofUsg) and the duration ratio of usage after changing to a new device out of the total months of usage (AfterdevchgP). Finally, we purposely categorized customers' churn motivation (ChurnMotive) which is target variable into major five variables by integrating values of small frequencies into one value, 'others'. In this way, the 12 variables were prepared for our experiments.

Table 1. Tweleve variables and available values for learning classifiers

Variable	Available values (number of values)
Age	10under, 10to12, 13to15, 16to19, 20to24, 25to29, 30to34, 35to39, 40to44, 45to49, 50to54, 55to59, 60over (13)
DeviceMaker	lge, motorola, pantech, samsung, skteletech, others (6)
ServiceGrade	BRONZE, GOLD, SILVER, VIP (4)
PaymentWay	AUTOBANK, CREDITCARD, JIRO (3)
CallsInaMon	-0.5, 0.5-2.5, 2.5- (3)
ARPUInaMon	-21, 21-3537, 3537-3599.5, 3599.5-9501, 9601-14810, 14810-33173.5, 33173.5-(7)
ARPUIn3MonAve	-2675, 2675-3567.5, 3567.5-4925.5, 4925.5-9444.5, 9444.5-15780.5, 15780.5-34978.5, 34978.5- (5)
VoiceFreqInaMon	-0.5, 0.5-85.5-, 85.5- (3)
VocieMinuInaMon	-7.5, 7.5-6854, 6854- (3)
*totMonofUsg**	*-37.5, 37.5- (2)*
*AfterdevchgP**	*-0.905, 0.905-, (2)*
ChurnMotive**	***failtopay, noneeds, numtrans, others(eg. Burden of high bill, stop to use of specific service (5)***

* manipulated by author, all other variables were left intact as donated.
** class node (target variable).

3.2 BN Classifiers

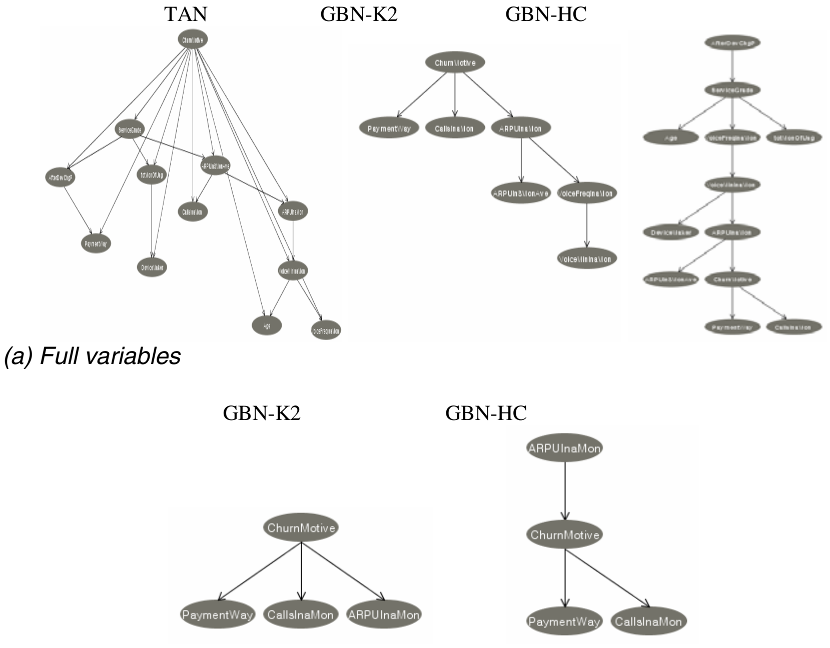
3.2.1 Structure Learning

We used WEKA [8], an open source data mining tool for various kinds of data mining, for BN constructs and experimental performance. The 11 variables (Table 2) were used to determine which networks had a target node of 'Churn motivation.' The structure of the GBN was learned using two search algorithms, K2 [5] and Hill Climber, with the maximum number of parent nodes limited to one. A BDeu scoring metric was used, a refinement of the K2 metric. The MDL scoring metric was used for Hill Climber (HC) [12]. For constructing NBN and TAN [6], the default setting in WEKA was used.

Table 2. Two types of variables

	Structure	NBN	TAN	GBN-K2	GBN-HC
Num. of nodes (Num. of Arc)	Full variables	12(11)	12(21)	7(6)	12(11)
	MB variables	-	-	4(3)	4(3)

Our experiments to create BN structures involved two types of variables, one for full variables and the other for MB variables. The later is purposed that only variables included in MB were used to create a structure, so the accuracy of classification was stable. Table 2 illustrates these types. The final analysis was conducted on 6 structures, which are illustrated in Figure 1. We purposely omitted NBN, because its shape is straightforward such that all the children nodes are directly linked with class node.



(a) Full variables

(b) Markov Blanket (MB) variables

Fig. 1. BN structures using two types of variables-full variables and MB variables

3.2.2 Results

(1) Prediction accuracy

For the sake of clear understanding of the experiment performance, we address the results from BN methods, and then the benchmarking classifiers' performance will be compared with BN performance.

Experiments with BN classifiers

First of all, let us state the results from experiments using full variables. Table 3(a) lists the performance of BN classifiers. Experiments were conducted with ten repetitions of ten-fold cross-validation on classification of mobile churn motivation. Prediction accuracy was measured in percent and figures in parenthesis indicate standard deviation. The best result is underlined in Table 3(a). To obtain rigor for each BN classifier's performance, each BN performance was compared with other BN classifiers' performance by using a corrected resample t-test at the 5% significance level based on the 10 x 10 cross-validation results. The results in Table 3(b) reassure that (1) GBN-HC and TAN show the statistically same performance, and (2) both GBN-K2 and NBN turn out to be equivalent statistically. In addition, comparing performance of BN classifiers, we came to understand that GBN-HC and TAN outperform GBN-K2 and NBN.

By the way, as stated previously, number of MB variables is considerably small compared with number of full variables. Furthermore, performance using MB variables is not bad in comparison with performance using full variables. This argument is confirmed by looking at the *Full variables vs MB variables* part of Table 3(b) where GBN-K2 performance using MB variables (47.68%) is better than GBN-K2 using full variables (47.16%), and GBN-HC using full variables shows statistically better performance than GBN-HC using MB variables (48.23% > 47.68%).

Judging from the discussion so far, it is clearly concluded that performance of BN classifiers using MB variables possess great potentials- (1) complexity measured by number of variables and arcs is very low in comparison with full variables, and (2) performance is not defeated seriously by the BN classifiers using full variables.

Table 3. Prediction Accuracies of BN Classifiers and Statistical Tests for Comparison

(a) Accuracies for each BN (unit: %)

Structure	NBN	TAN	GBN-K2	GBN-HC
Full variables	46.78(1.62)	48.27(1.58)	47.16(1.83)	48.23(1.72)
MB variables	-	-	47.68(1.72)	47.68(1.73)

(b) t-test results for BN classifiers

	Paired Differences					t-value	Sig. (2-tailed)
				95% Confidence Interval of the Diff.			
	Mean	Std. Dev.	Std. Error Mean	Lower	Upper		
Full variables							
NBN - TAN**	-1.49	1.99	0.199	-1.89	-1.09	-7.49	0.00
NBN – GBN K2	-0.38	2.18	0.218	-0.82	0.05	-1.76	**0.08**
NBN - HC**	-1.46	2.40	0.240	-1.94	-0.98	-6.07	0.00
TAN - K2**	1.11	2.42	0.242	0.63	1.59	4.58	0.00
TAN – GBN HC	0.03	2.36	0.236	-0.44	0.50	0.14	**0.89**
K2 - HC**	-1.07	2.23	0.223	-1.52	-0.63	-4.82	0.00
Full variables vs MB variables							
Full GBN K2 -MB GBN K2	-0.52	1.25	0.125	-0.77	-0.28	-4.21	0.00
Full GBN HC- MB GBN HC	0.56	0.69	0.069	0.42	0.69	8.07	0.00

$* p < 0.05, ** p < 0.01.$

Experiments with benchmarking classifiers

Benchmarking classifiers are necessary to verify the validity of accuracies of our proposed BN approach in Table 4. They include classical types of classifiers such as SVM (Support Vector Machine), Neural network (NN), and Decision tree. For the sake of SVM, LibSVM[2] was used. DT was C4.5 [15] and its module was adopted from *J48* of WEKA. NN used for benchmarking classifier was a *MultilayerPerceptron* module supported by WEKA. For the sake of computational rigor, 10-fold cross-validation was performed 10 times, and average performance was obtained for each classifier. To show that the GBN-based approach predicting mobile churn motivations outperforms benchmarking classifiers and other types of BNs, we performed paired-samples t-test

[2] `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

using SPSS 12.0. Table 4 summarizes prediction accuracies of benchmarking classifiers. Test results show that prediction accuracies of SVM, C4.5, TAN and GBN-HC are not different statistically from each other.

Table 4. Benchmarking Classifiers and Its Comparison with BN classifiers

(a) Performance of benchmarking classifiers (unit: %)

Classifiers	SVM	NN	J48
performance	42.82(1.61)	43.92(2.30)	47.96(1.82)

Figures in parenthesis indicate standard deviation.

(b) t-test results for benchmarking classifiers

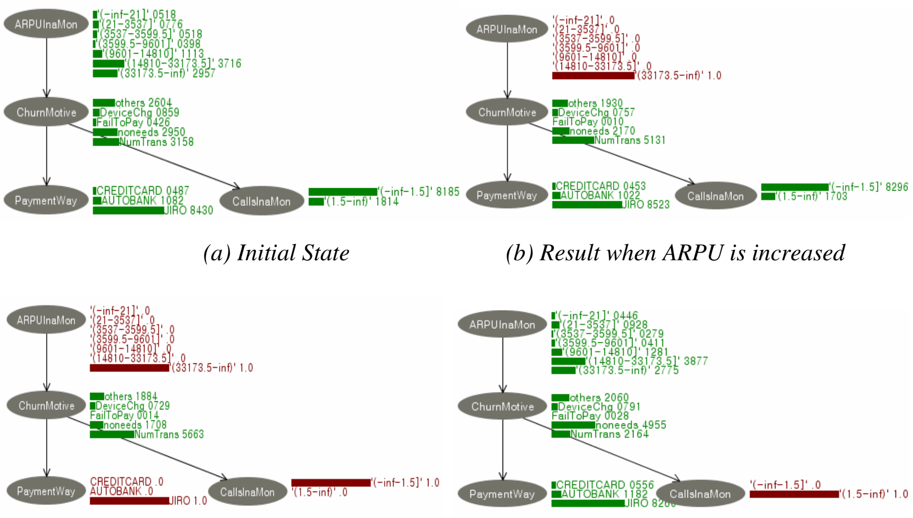
	Paired Differences						
				95% Confidence Interval of the Diff.		t-value	Sig. (2-tailed)
	Mean	Std. Dev.	Std. Error Mean	Lower	Upper		
SVM - NN**	3.89	2.46	0.246	3.41	4.38	15.82	0.00
SVM - J48	-0.14	2.36	0.235	-0.61	0.32	-0.61	**0.54**
SVM - NBN**	1.04	2.19	0.216	0.61	1.48	4.74	0.00
SVM - TAN	-0.45	2.39	0.239	-0.92	0.02	-1.88	**0.06**
SVM - K2**	0.66	2.42	0.242	0.18	1.14	2.71	0.01
SVM - HC	-0.42	2.33	0.233	-0.88	0.04	-1.79	**0.08**
NN - J48**	-4.04	2.85	0.285	-4.60	-3.47	-14.17	0.00
NN - NBN**	-2.85	2.86	0.286	-3.42	-2.29	-9.99	0.00
NN - TAN**	-4.34	2.83	0.283	-4.91	-3.78	-15.37	0.00
NN - K2**	-3.24	3.06	0.306	-3.85	-2.63	-10.57	0.00
NN - HC**	-4.31	2.84	0.284	-4.88	-3.74	-15.18	0.00
J48 - NBN**	1.18	2.41	0.241	0.71	1.66	4.92	0.00
J48 - TAN	-0.31	2.49	0.249	-0.80	0.187	-1.23	**0.22**
J48 - K2**	0.80	2.55	0.255	0.29	1.31	3.13	0.00
J48 - HC	-0.27	2.55	0.255	-0.78	0.23	-1.07	**0.29**

* $p < 0.05$, ** $p < 0.01$.

Then, at this moment, we have to raise two criteria by which method can be compared with each other fairly. First criterion is whether causal relationships can be induced from the method. Values of causal relationships should be measured by the characteristics of our target problem- prediction of mobile churn motivations. For the mobile telecommunication service providers investigating motivations of users to churn, such causal relationships are essential. In addition, number of churning motivations are numerous, indicating that company should spend a lot of time and efforts to pin down exact reasons of churning and amend operational and/or marketing pitfalls to prevent further churning. Second criterion is whether the number of variables can be reduced logically without loss of prediction accuracy. From this perspective, the GBN is very good at reducing the number of variables to be considered because its MB properties guarantee such possibility. This means telecommunication service providers are able to predict churn motivation more efficiently with compacter form of causal relations among MB variables. Thus, we continue to analyze practical implication with BN classifiers using MB variables in the following section.

(2) What-if analysis

The motivation of churn can be predicted by performing what-if analyses with GBN structure. Let us investigate what-if results with GBN-HC structure using MB variable. Mobile telecommunications service providers want to know how many customers are likely to transfer to competing companies, which is represented by the 'NumTrans' value in the class node 'ChurnMotive'. Figure 2 illustrates the what-if analysis results showing that customers are more likely to transfer to a competing company when (1) their ARPU of the previous month was rather high (2) they pay their bill using JIRO and (3) they call contact center very few.



(a) Initial State

(b) Result when ARPU is increased

(c) Result when payment method is JIRO and there are few calls for contact center

(d) Result when there are few calls for contact center

Fig. 2. What-if analyses using MB variables

3.3 Discussion

First of all, it should be noted that GBN classifiers can provide a set of causal relationships among relevant variables with target node, and then decision makers are able to find useful strategies to prevent undesirable churning behavior by forecasting which kinds of customers are believed to have churning motivation. From the performance perspective, GBN classifiers are showing competitive prediction accuracies compared with other BN classifiers. Therefore, we found usefulness of GBN classifiers in the problem area of predicting customers' churning motivation.

On the other hand, value of MB variables must be mentioned here. MB variables are those variables provided by only GBN classifiers. What is striking with the MB variables is that prediction accuracies by GBN with MB variables are very competitive compared with other BN classifiers using full variables. Therefore, using the MB variables can provide a number of advantages for decision makers who seek more compact decision mechanism where number of decision variables to be considered by decision makers should be kept to minimum.

Lastly, GBN classifiers are capable of uniquely providing what-if simulation functions with which decision makers can test various numbers of alternative solutions to the target problem. In case of this paper, decision makers can perform a lot of what-if simulations before deciding a final strategy to prevent customers' churning behavior which is undesirable to companies' profitability.

4 Concluding Remarks

Previous studies about mobile churn prediction have always handled churning behavior itself. However, contrary to this research trend, this study is aimed at predicting churn motivations by using BN classifiers and then comparing their performance with benchmarking classifiers. Let us summarize our contributions as follows.

First, four types of BN classifiers were considered- NBN, TAN, GBN-K2, and GBN-HC. Besides, to show validity of MB variables provided by GBN, we tested the performance by MB variables with other BN classifiers. From this attempt, usefulness of using GBN assisted by MB variables is very high, especially in the field of decision problems where a lot of decision variables should be considered before suggesting a final solution.

Second, flexible properties of what-if simulation must also be highlighted. Such what-if simulation capability is found only in BN classifiers. Therefore, BN classifiers are recommendable to be used in resolving complicated decision problems like churning motivation prediction, once their prediction accuracy is competitive.

Further study issues remain. For example, ensemble approach seems necessary to improve prediction accuracy. Another type of future study issue is to determine relevant number of variables to describe churning motivation.

Acknowledgments. This research is supported by the Ubiquitous Computing and Network (UCN) Project, Knowledge and Economy Frontier R&D Program of the Ministry of Knowledge Economy (MKE) in Korea as a result of UCN's subproject 09C1-T2-20S.

References

1. Ahn, J., Han, S., Lee, Y.: Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecommunications Policy 30(10-11), 552–568 (2006)
2. Bouckaert, R.: Properties of Bayesian belief network learning algorithms. In: Proc. 10th Annual Conf. Uncertainty Artificial Intelligence (UAI), Seattle, WA, pp. 102–110 (1994)
3. Chen, P.-Y., Hitt, L.M.: Measuring switching costs and the determinants of customer retention in Internet-enabled businesses: Astudy of the online brokerage industry. Information Systems Research 13(3), 255–274 (2002)
4. Cheng, J., Greiner, R.: Comparing Bayesian network classifiers. In: Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence, pp. 101–107. Morgan Kaufmann Publishers, San Francisco (1999)

5. Cooper, G.F., Herskovits, E.: A Bayesian method for the induction of probabilistic networks from data. Machine Learning 9(4), 309–347 (1992)
6. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine Learning 29(2), 131–163 (1997)
7. Gemela, J.: Financial Analysis Using Bayesian Networks. Applied Stochastic Models in Business and Industry 17, 57–67 (2001)
8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter 11(1), 10–18 (2009)
9. Kim, M.K., Jeong, D.H.: The effects of customer satisfaction and switching barriers on customer loyalty in Korean mobile telecommunication services. Telecommunications Policy 28(2), 145–159 (2004)
10. Koller, D., Sahami, M.: Toward optimal feature selection. In: Proc. 13th International Conf. Machine Learning, pp. 284–292 (1996)
11. Lemmens, A., Croux, C.: Bagging and Boosting Classification Trees to Predict Churn. Journal of Marketing Research 43(2), 276–286 (2006)
12. Madden, M.G.: On the classification performance of TAN and general Bayesian networks. Knowledge-Based Systems 22(7), 489–495 (2009)
13. Neslin, S.A., Gupta, S., Kamakura, W., Lu, J., Mason, C.: Defection detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing Research 43(2), 204–211 (2006)
14. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco (1988)
15. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993)
16. Sarkar, S., Sriram, R.S.: Bayesian Models for Early Warning of Bank (1998)
17. Siber, R.: Combating the churn phenomenon. Telecommunications 31(10), 77–81 (1997)
18. Yan, L., Wolniewicz, R.H., Dodier, R.: Predicting customer behavior in telecommunications. IEEE Intelligent Systems 19(2), 50–58 (2004)
19. Wei, C.P., Chiu, I.T.: Turning Telecommunications Call Details to Churn Prediction: A Data Mining Approach. Expert Systems with Application 23, 103–112 (2002)

Logical Mechanism for Allocating Resources to Exploitation and Exploration to Achieve Ambidexterity: Team Level Analysis

Kun Chang Lee[1] and Do Young Choi[2,*]

[1] Professor of MIS and WCU Professor of Creativity Science, SKK Business School and
Department of Interaction Science, Sungkyunkwan University, Seoul 110-745,
Republic of Korea
Tel.: +82 2 760 0505; Fax.: +82 2 760 0440
kunchanglee@naver.com, leekc@skku.edu
[2] Principal Consultant, LG CNS, Seoul 100-725, Republic of Korea
dychoi96@gmail.com, choidy@lgcns.com

Abstract. Despite the rise of studies on knowledge creation in the perspective of exploitation and exploration in organizational learning, previous studies are rare which suggested a concrete mechanism regarding how to allocate limited resources to exploitation and exploration to remain ambidextrous. Main purposes of this paper are to make logical argument on how teams create creativity through knowledge creation by balancing exploitation and exploration, and to present a new logical mechanism by which teams allocate their limited resources to exploitation and exploration to achieve balance between them. Time-dependent simulations were conducted to prove the validity of the proposed logical mechanism for sustaining the balance between exploitation and exploration.

Keywords: Team creativity, Exploitation, Exploration, Ambidexterity, Logical mechanism, Knowledge creation.

1 Introduction

Studies on organizational learning and knowledge creation have covered exploration which develops new solutions by searching for new knowledge and exploitation which finds new solutions by utilizing existing knowledge [11, 15, 17]. These studies stress that organizations should pursue both exploitation and exploration for sustaining competitive capability and it is essential for organization to have balanced resource allocation between exploitation and exploration for organizational survival and prosperity because resources are limited in organizations [5, 12, 15].

Although the previous studies on exploitation and exploration have consensus that exploitation and exploration are important for organizations' success and balance between them should be achieved [7, 9, 13, 15], the studies on the concrete mechanism of balancing between exploitation and exploration are scarce [9]. In the most of the

[*] Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 314–323, 2010.

studies on the balancing between exploitation and exploration, ambidexterity [1, 2, 3, 12, 16] – ambidextrous organization – and punctuated equilibrium [4] are considered as organizational method that can achieve the balance between exploitation and exploration. These kinds of researches addressed the importance and the balance in the macro-level perspectives – whole organization level or inter-organizational level. Therefore, the studies on exploitation and exploration in the micro-level - individual level or team level – are scarce [9], and scarce are the studies on concrete criteria and mechanism of how to achieve balance between exploitation and exploration. Therefore, in this paper we address the knowledge creation structure as creativity revelation process based on exploitation and exploration in the perspective of ambidexterity with mathematical modeling, and we present resources allocation criteria and mechanism within a team as explanation for balance between exploitation and exploration.

The purpose of this study is to make logical argument on how organizations allocate their limited resources between exploitation and exploration to achieve balance between them. As stated previously, studies of organizational theorists and strategic management theorists have stressed the strategic importance of exploitation/exploration choice and the importance of resource balancing between them in order to maintain organizations' success and competitive capability [5, 12, 15]. In order to find the appropriate methods to achieve balance, they mostly focus on organizational design or behavioral and social means. However, there seems to be few researches to present established mechanism or concrete criteria to allocate limited resources between exploitation and exploration to achieve balance. That is, previous studies based on organizational theory or strategic management theory have focused on the strategic necessity of exploitation and exploration and on the balancing mechanism in the perspective of organizational architecture and management method in the macro level – organizational level or inter-organizational level [9]. Therefore, this paper focuses on team level analysis to make concrete criteria and direct mechanism to allocate resources between exploitation and exploration as balancing mechanism.

This paper is organized as follows. First, we conduct literature review in two perspectives related to the purpose of this study; 1) Relationship and balancing between exploitation and exploration in strategy and organizational perspective, 2) Balancing mechanism between exploitation and exploration in analytical perspective including simulation method. Second, we present the knowledge creation model theoretically based on creativity revelation process of through balancing between exploitation and exploration. This model explains how team knowledge can be created through balancing mechanism between exploitation and exploration which are important processes to reveal team creativity. Evolution pattern of team knowledge creation is presented over time. Finally, we discuss the implications of our model for management and further research.

2 Literature Review

The studies on exploitation and exploration based on organizational theory and strategic management theory have covered the strategic importance and organizational choice of them for organization's short-term and long-term success [5, 12, 15]. They also addressed the managing mechanism of balance between exploitation and exploration in the viewpoint of organizational structure or in the behavioral viewpoint [1, 2, 16], and they examined the related performance empirically [10, 13-14]. Since March's study [15] on the adaptation mechanism between exploitation and exploration, many researchers addressed this topic in the various perspectives; organizational structure and management behavior [1, 2, 16], empirical test based on field researches [10, 13-14], and mathematical simulation modeling in the perspective of solution finding and knowledge creation processes [6, 8, 11].

2.1 Strategy and Organizational Perspective

From the perspective of organization structure and resource allocation, strategic importance of how to balance between exploitation and exploration was extensively discussed [15, 18]. Though definitions about exploitation and exploration are still debatable depending on researchers, there exists an agreement among researchers that how to balance between exploitation and exploration is extremely crucial for the sake of organizational performance, short-term or long-term [9]. Among many theories about this important issue, i.e. *how to balance between exploitation and exploration*, most popular approach is to seek such balance through *ambidexterity* or *punctuated equilibrium*. This approach is based on a notion that organizational structure is an important vehicle for balancing between exploitation and exploration. For example, balancing through ambidexterity is to seek such balance through building loosely coupled and differentiated subunits or individuals.

There exist two kinds of approaches to determining the balance between exploitation and exploration. First one is to balance them through ambidexterity, in which the maximum performance should be sought in both exploration-specific subunits and exploitation-oriented subunits without partially placing emphasis on one of them. In contrast, punctuated equilibrium approach is that initial emphasis should be given to one of exploration and exploitation, and then next priority is sequentially given to other after a certain level of performance is accomplished. Therefore, temporal cycle between exploration-focus and exploitation-focus is inevitable in the punctuated equilibrium approach. Raisch et al. [16] studied the three perspectives related to accomplishing the organizational ambidexterity aiming balance between exploitation and exploration for the sake of sustainable organizational performance- how to obtain ambidexterity (differentiation or integration), level of pursuing ambidexterity (individual level or organizational level), and temporal perspective (static or dynamic). As He and Wong [10] noted, empirical studies emphasizing to answer the research question about the effect of ambidexterity on the firm performance are rare in literature. In that sense, He and Wong [10] showed through empirical approach that ambidextrous organizations placing emphasis on technological innovation yield better performance than punctuated organizations. In other words, ambidextrous firms show strong trend in sales growth, proving conventional arguments that that the organizational balancing between exploitation and exploration is essential for the organizational success. Andriopoulos and Lewis [1] proposed two kinds of materializing ambidexterity. First one is architectural ambidexterity focusing on the use of organizational structure and strategy to enable proper differentiation of exploitation and exploration. Second one is contextual ambidexterity utilizing more behavioral and social means to integrate exploitation and exploration. From a single case of a firm which has centered only on exploitation for growth, and almost faced bankruptcy, McNarama and Baden-Fuller [14] suggested that it is possible for such a firm to renew its growth engine based on exploration and stay on the balance of exploration and exploitation. Liu [13] showed that the imbalancing (i.e., excessive exploration or excessive exploitation) leads to competency trap, degrading firm performance. Accordingly, he stressed the importance of organizational ambidexterity.

2.2 Analytical Perspective

Analytical perspective introduces either mathematical modeling or agent-based modeling to obtain an appropriate way of balancing the exploitation and exploration. March [15] launched a pioneering study to show how important it is for firms to maintain the balance of exploitation and exploration in order to remain competitive in the market. Following March [15]'s exploration-exploitation model, various kinds of analytical methods were proposed for the sake of balancing the exploitation and exploration [6, 8]. For example, by referring to simulation method Fang et al. [6] suggested the usefulness of maintaining semi-isolated subgroups in organization and varying the interaction pattern between individuals to obtain the balance of exploitation and exploration. With the aid of system dynamics approach, Garcia et al.[8] identified importance of the four factors in balancing the exploitation and exploration in R&D activities in technology-oriented companies- resource availability, exogenous competition, aging of knowledge based, and adaptive capacity.

3 Team Creativity Model

We assume that team creativity is increased by knowledge creation supported by balancing exploitation and exploration. Resources are scarce and should be distributed very carefully to sustain ambidextrous. Ambidexterity is achieved by balancing exploitation and exploration. It seems that deliberate resource allocation mechanism among exploitation and exploration is crucial for achieving such ambidexterity. Considering research questions like this, research model is proposed as shown in Figure 1. Definitions of exploitation and exploration are different among researchers [6]. We adopt the definition by March [15], and Lazer and Friedman [11]. Therefore, exploitation is defined as knowledge enhancement or creation by utilizing of known knowledge and known solutions. Also exploration is defined as new knowledge creation by development of unknown knowledge and unknown solutions.

Our proposed research model is depicted in Figure 1 where knowledge creation occurs by exploitation and exploration, and then team creativity increases accordingly. For teams to increase their creativity level, they seem to need more knowledge. Here we assume that such knowledge should be created through balance between exploitation and exploration. In other words, teams are supposed to seek short-term performance improvement by allocating resources to exploitation, and pursue long-term performance by allocating some portion of resources to exploration activities. However, teams should be very cautious not to allocate their limited resources too much to either exploitation or exploration, losing its balance. In a dynamic market condition, teams are strongly monitored to achieve such balance between exploitation and exploration so that both short-term performance and long-term growth can be obtained simultaneously.

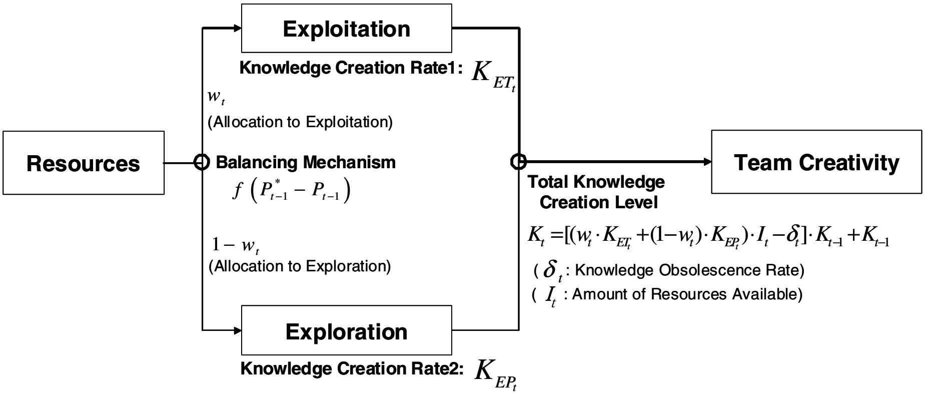

Exploitation
Knowledge Creation Rate1: K_{ET_t}
w_t
(Allocation to Exploitation)
Resources
Balancing Mechanism
$f\left(P^*_{t-1} - P_{t-1}\right)$
$1 - w_t$
(Allocation to Exploration)
Exploration
Knowledge Creation Rate2: K_{EP_t}
Team Creativity
Total Knowledge Creation Level
$K_t = [(w_t \cdot K_{ET_t} + (1-w_t) \cdot K_{EP_t}) \cdot I_t - \delta_t] \cdot K_{t-1} + K_{t-1}$
(δ_t : Knowledge Obsolescence Rate)
(I_t : Amount of Resources Available)

Fig. 1. Team Creativity Revelation Model

3.1 Resource Allocation between Exploitation and Exploration

Exploitation and exploration compete for limited resources in organization [15]. So teams should have appropriate criteria for allocating resources to each of them. Essentially teams are evaluated based on their performances compared to their given goals. Assuming that team performance increases as team creativity is facilitated more, we need to handle resource allocation sophisticatedly to maximize effects of exploitation and exploration. As mentioned earlier, exploitation aims short-term performance and exploration looks for long-term performance relatively compared with exploitation. Therefore, teams should allocate their resources to keep balance between exploitation and exploration, which is termed “ambidexterity”. Now, we suggest a related proposition.

***Proposition**: Limited resources are allocated to exploitation and to exploration by the gap between team’s target performances and its actual performances. In other words, the team with over-achievement can allocate more resources to exploration with long-term view. In contrast, the team with under-achievement should allocate more resources to exploitation with short-term view for the fast improvement of performance.*

Since effects of exploitation and exploration can be observed with time, the proposed resource allocation mechanism is introduced in time *t*. Suppose that decision makers pay attention only to the discrepancy between actual performance and target performance. If actual performance exceeds target performance (i.e., $P^*_{t-1} - P_{t-1} < 0$), more resources should be allocated to exploration at time *t* to obtain long-term growth capability. On the other hand, if actual performance falls short of target performance (i.e., $P^*_{t-1} - P_{t-1} > 0$), then more resources must be allocated to exploitation in order for short-term performance to be achieved quickly. Therefore, the proposed resource allocation function at time *t* is given by the following equation (1):

$$w_t = f(P^*_{t\text{-}1} - P_{t\text{-}1}) = \frac{1}{1+e^{-c(P^*_{t\text{-}1} - P_{t\text{-}1} + g)}} \tag{1}$$

where $0 \le w_t \le 1$, w_t refers to the resources allocation weight to exploitation at time *t*, P^*_{t-1} the target performance at time *(t-1)*, and P_{t-1} the actual performance at time *(t-1)*. Inversely, (1- w_t) means the resources allocation weight to exploration at time *t*. By adjusting function (1), how to allocate limited resources to exploitation will be determined. Such implications are depending on the interpretation of *c* and *g* used in function (1) above.

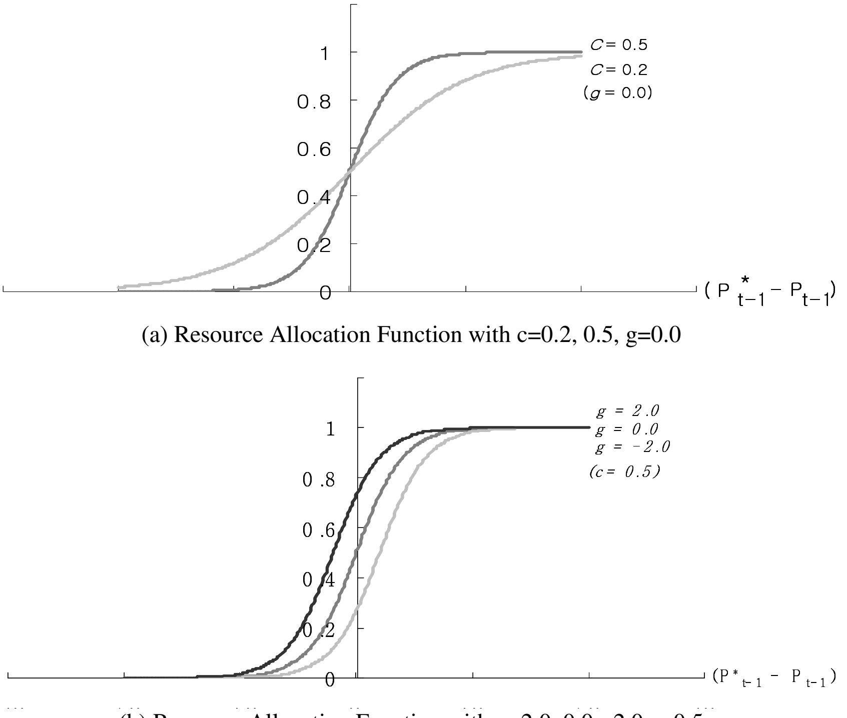
Implications of constant "c"

Constant *c* implies how much team is sensitive to its performance gap. If a team shows a big *c*, it means that the team is very sensitive to poor performance. Those teams related to sales and manufacturing are usually showing a big *c*, and therefore they will try to allocate a big chunk of resources to exploitation for the sake of improving performance fast when their performance is poor than expected. Meanwhile, teams with a small *c* indicate that they are rather insensitive to poor performance. Henceforth, in case of poor performance, such teams like general management, accounting, and human resources management are likely to try to allocate some resources to exploitation for boosting their performance. However, for the teams with small *c*, the amount of resources allocated to exploitation is quite smaller than that of teams with big *c*.

Implications of constant "g"

g shows a reference point of resource allocation to exploitation and exploration. The reference point varies with team mission. For example, some teams are supposed to focus on exploration rather than exploitation, and vice versa. Typically, R&D teams are supposed to put their primary efforts on long-term outcomes, and therefore they should be bent for allocating their resources to exploration more than to exploitation. In contrast, sales teams are concerned with short-term performance, pushing them to allocate more resources to exploitation. For those teams aiming exploitation, g must have negative values to shift resource allocation curve to right direction. On the contrary, positive values of g are assigned to the teams pursuing exploration. Therefore, positive g values are suitable for R&D teams, and negative g values are appropriate for sales teams.

Figure 2 depicts different shapes of the resource allocation function depending on *c* and *g*. Considering the context which teams have, the resource allocation criteria can be moved by controlling constant number *g* (refer to the Fig. 2). This function is shaped as follows:



(a) Resource Allocation Function with c=0.2, 0.5, g=0.0

(b) Resource Allocation Function with g=2.0, 0.0, -2.0 c=0.5

Fig. 2. Different Shapes of Resource Allocation Functions

3.2 Total Knowledge Creation Level

Team creativity is determined by team knowledge creation through balancing exploitation and exploration. Also, team knowledge creation is affected by several factors such as team environment, task characteristics, and interactions among team members [12, 17]. As previously described in our research model – team creativity revelation model, team knowledge creation occurs by exploitation and exploration itself. Each exploitation and exploration process has its own unique knowledge creation process by utilizing available resources provided by team.

Accordingly, K_{ET} and K_{EP} represent how much knowledge can be created on the basis of resources allocated to exploitation and exploration, respectively. Meanwhile, some portion of existing knowledge becomes obsolete as new knowledge is created over time and accumulated inside team. Therefore, we have total knowledge level function at time *t* as shown in the following equation (2):

$$K_t = [(w_t \cdot K_{ET_t} + (1 - w_t) \cdot K_{EP_t}) \cdot I_t - \delta_t] \cdot K_{t-1} + K_{t-1} \tag{2}$$

where K_t refers to total knowledge level at time *t*, w_t the resources allocation weight to exploitation at time *t*, K_{ET_t} the knowledge creation rate from exploitation at time *t*,

K_{EP_t} the knowledge creation rate from exploration at time *t*, I_t the amount of resources available at time *t*, δ_t knowledge obsolescence rate at time *t*, and K_{t-1} total knowledge level at time *(t-1)*.

Considering the definition of exploitation and exploration, it is easily expected that knowledge creation patterns from exploitation and exploration would be different from each other. Let us guess at this moment how the pattern would be. Through exploitation process, team creates knowledge by utilizing the existing knowledge and solutions, which makes us aware that teams accumulate knowledge incrementally by exploitation process. Meanwhile, exploration focuses on new knowledge creation by development of unknown knowledge and solutions. Therefore, through exploration process, team creates a new knowledge at a quantum jump rate at certain time rather than incremental type of creation.

By keeping balance between exploitation and exploration, simulation was conducted over we conducted the simulation of total knowledge creation level over 60 time intervals. Knowledge creation pattern is depicted in Figure 3, where team creativity is represented by the level of cumulative knowledge created. At around time interval = 40, team creativity level jumped with a great momentum, indicating that exploration efforts were successful at that time, and consequently knowledge creation level increased drastically.

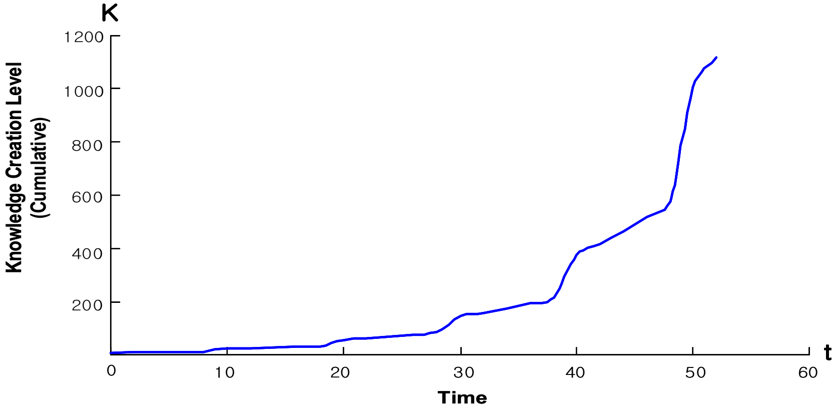


Fig. 3. Total Knowledge Creation Level: Time Dimension

4 Implications and Limitations

As mentioned so far, how team creativity is created in organizations was analyzed with an emphasis on balancing between exploration and exploitation. Though similar researches about ambidexterity were performed, this study is unique in the sense that logical structure is proposed about how to balance the exploration and exploitation to remain ambidextrous, and the team creativity revelation process is also suggested with the proposed logical structure. Results of this study are very significant from the perspective that dynamic resource allocation mechanism is firstly proposed for companies to remain ambidextrous by balancing between exploitation and exploration. The implications of this study are as follows.

First, results of this study can be used as a theoretical foundation for agent-based modeling study. The proposed logical mechanism about how to remain ambidextrous can support further studies related to those studies using agent-based model (ABM).

Second, in the perspective of organizational structure and managerial behavior theory this paper can expand the argument about resources allocation and balancing between exploitation and exploration.

Limitations also exist. First, basic assumption should be relaxed in the further study. Our basic assumption was that exploitation and exploration are regarded as exogenous factors and generated by its own process[1]. Further study should relax this assumption to suggest more realistic knowledge creation mechanism that is used to describe how team creativity is yielded accordingly by adjusting exploitation and exploration themselves. Second, the proposed knowledge creation mechanism should be empirically proved by either survey method or ABM. In authors' view, ABM approach would yield more robust and meaningful results for both practitioners and academicians. Third, knowledge creation process by both exploitation and exploration should include a set of appropriate relevant factors such as individual knowledge level, task complexity, and several social network metrics among team members (i.e., degree centrality, between-ness centrality, heterogeneity, and structural hole, etc)

5 Conclusion

Strategy and organization theory emphasizes the importance of balancing exploitation and exploration for the sake of organizations' sustainable success. Even though there exist approaches like ambidexterity and punctuated equilibrium that were suggested as methods to achieve this balance, there is no study to attempt to propose a concrete mechanism about how to balance between exploration and exploitation. This study proposed a logical mechanism in which a dynamic resource allocation mechanism plays to balance exploitation and exploration, and then team creativity is based on such knowledge creation facilitated by the balancing activity. In this way, we revealed successfully with a rigorous logic that how team creativity is organized through knowledge creation supported by balance of exploitation and exploration. Our results are directly related to a very important issue of how an organization can remain ambidextrous.

Acknowledgment

This work was supported by the National Research Foundation Grant funded by the Korean Government (342-2009-1-B00015).

References

1. Andriopoulos, C., Lewis, M.W.: Exploitation-exploration tensions and organizational ambidexterity: Managing paradoxes of innovation. Organization Science 20(4), 696–717 (2009)

[1] This process is not described in this paper to save space. However, this process should be modeled more elaborately by using an appropriate theory.

2. Benner, M.J., Tushman, M.L.: Exploitation, exploration, and process management: The productivity dilemma revisited. Academy of Management Review 28(2), 238–256 (2003)
3. Burgelman, R.A.: Intra-organizational ecology of strategy-making and organizational adaptation. Organization Science 2, 239–262 (1991)
4. Burgelman, R.A.: Strategy as vector and the inertia of coevolutionary lock-in. Administrative Science Quarterly 47, 325–357 (2002)
5. Cohen, W.M., Levinthal, D.A.: Absorptive capacity: A new perspective on learning and innovation. Administrative Science Quarterly 35, 128–152 (1990)
6. Fang, C., Lee, J., Schilling, M.A.: Balancing exploration and exploitation through structural design: The isolation of subgroups and organization learning. Organization Science 21(3), 625–642 (2010)
7. Feinberg, S.E., Gupta, A.K.: Knowledge spillovers and the assignment of R&D responsibilities to foreign subsidiaries. Strategic Management Journal 25, 823–845 (2004)
8. Garcia, R., Calantone, R., Levine, R.: The role of knowledge in resource allocation to exploration versus exploitation in technologically oriented organizations. Decision Sciences 34(2), 323–349 (2003)
9. Gupta, A.K., Smith, K.G., Shalley, C.E.: The interplay between exploration and exploitation. Academy of Management Journal 49(4), 693–706 (2006)
10. He, Z., Wong, P.: Exploration vs. exploitation: An empirical test of the ambidexterity hypothesis. Organization Science 15(4), 481–494 (2004)
11. Lazer, D., Friedman, A.: The network structure of exploration and exploitation. Administrative Science Quarterly 52, 667–694 (2007)
12. Levinthal, D.A.: Adaptation on rugged landscapes. Management Science 43, 934–950 (1997)
13. Liu, W.: Knowledge exploitation, knowledge exploration, and competency trap. Knowledge and Process Management 13(3), 144–161 (2006)
14. McNamara, P., Baden-Fuller, C.: Lessons from the Celltech case: Balancing knowledge exploration and exploitation in organizational renewal. British Journal of Management 10, 291–307 (1999)
15. March, J.G.: Exploration and exploitation in organizational learning. Organization Science 2(1), 71–87 (1991)
16. Raisch, S., Birkinshaw, J., Probst, G., Tushman, M.L.: Organizational ambidexterity: Balancing exploitation and exploration for sustained performance. Organization Science 20(4), 685–695 (2009)
17. Reagans, R., Zuckerman, E.W.: Networks, diversity, and productivity: The social capital of corporate R&D teams. Organization Science 12(4), 502–517 (2001)
18. Teece, D.J., Pisano, G., Shuen, A.: Dynamic capabilities and strategic management. Strategic Management Journal 18, 509–533 (1997)

Analyzing Economic Impact of Disruptive Technology Using Multi-Agent Simulation: Smart Payment Case

Kun Chang Lee[1], Young Wook Seo[2,*], and Min Hee Hahn[3]

[1] Professor of MIS at SKK Business School,
WCU Professor of Creativity Science at Department of Interaction Science,
Sungkyunkwan University,
Seoul 110-745, Republic of Korea
kunchanglee@gmail.com, leekc@skku.edu
[2] Principal Researcher,
Software Engineering Center at NIPA(National IT Industry Promotion Agency),
Seoul 138-711, Republic of Korea
Tel.: +82 2 760 0505; Fax: +82 2 760 0440
seoyy123@gmail.com
[3] Researcher,
Business Management Unit, LG CNS CO., Ltd., Republic of Korea
minheehahn@gmail.com, hahnminhee@skku.edu

Abstract. Disruptive technology creates disruptive impacts, although it takes time to identify radical technological change and analyze its subsequent economic impacts in the industry. Despite the characteristics of disruptive technology, empirical research in this area has focused on case studies and has not attempted time-variant simulation to investigate its long-time effects. To address this research void, this study adopts a multi-agent simulation technique to analyze long-time effects of a smart payment method which is regarded as a disruptive technology. Experimental results via the multi-agent simulation are meaningful and robust, and their practical implications are discussed.

Keywords: Disruptive Technology, Smart Payment, Mobile Payment, Traditional Payment, Multi-Agent Simulation.

1 Introduction

The recent advent of Web technology changes many aspects of our daily life, finances in particular. Mobile banking is now common, with mobile payments, defined as payment via mobile devices such as a cellular phone or smart phone [19, 22], being more typical than in-person bank visits. For the sake of clarity, we assume that mobile payment is based on general types of mobile phones and devices, and smart payment is used only on smart phones. Therefore, when "mobile payment" is mentioned, we are excluding that made via smart phones.

[*] Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 324–333, 2010.

This paper deals with the economic analysis of disruptive techniques, such as smart payment. A disruptive technology is a radical and disruptive technology system that neutralizes the existing technologies or market thresholds by offering completely new functions and properties [1, 5, 6, 8].

Smart payment has the potential of becoming a disruptive technique for the following reasons. First, smart payment is immediately accessible. This technique is, therefore, more readily accessed than is a credit card.

Second, smart payment looks "cool" to young generations and to the many consumers who seek individuality. Using mobile payments will allow for increased freedom, providing a sense of constant connectivity.

Third, smart payment provides more convenience than do credit cards. Smart phones are capable of providing several features of smart payment, such as real-time support of segmented and personalized shopping information, credit payment, mileage accumulation, and integration of advertising and coupons.

Despite these advantages of smart payment, it is necessary to consider additional fees that occur for parties using the smart payment method. When smart payment is adopted as a widely accepted payment method in the market, consumers are supposed to pay fees for its use, as do merchants, telecommunication companies, and financial institutions.

However, market analysis of the adoption of disruptive technology is a highly ill-structured problem that has not determined the efficiencies of the conventional approaches [5, 7]. Therefore, this study adopts a heuristic approach using a multi-agent simulation (MAS), and then applies it to the resolution of the research question: *"How does a disruptive technology like smart payment influence interested parties over time?"*

The main objectives of this study are as follows. First, interested parties are those market players who are going to actively use the method in question. In this study, there exist four types of interested parties: customers, merchants, telecommunication companies, and financial institutions. Second, the effects of smart payment on those four players are analyzed using longitudinal MAS.

2 Theoretical Background

2.1 Disruptive Technology

The term "disruptive technology" was first coined by Christensen and Bower [2], who further developed this concept in the 1997 book, *The Innovator's Dilemma*, in which the term "disruptive technology" was transformed into "disruptive innovation".

Disruptive technology involves a radical and disruptive technology system that neutralizes the existing technologies or market thresholds by offering completely new functions and properties.

As proposed by Adner [1], Charitou and Markides [5], Christensen [8], Christensen and Bower [6], Christensen and Raynor [7] and Gilbert [9], disruptive technology takes on the following five characteristics. First, the innovation underperforms with respect to the attributes valued by mainstream customers. Second, new functions that are provided by disruptive technology are not highly appreciated by mainstream customers. Third, disruptive technology is much simpler and less expensive, conventionally speaking, and it is offered at a lower price than are the current products. Fourth, at the time of its introduction, the innovation appeals to a low-end, price-sensitive customer segment, thus limiting the profit potentials for incumbents. Fifth, over time, further developments improve the innovation's performance with regard to the valued attributes of mainstream customers to such a degree that the innovation begins to attract more of these customers.

Ondrus and Pigneur [16] looked into disruptions of mobile payment from two perspectives, the move from credit card to mobile phone and the replacement of manager-oriented solutions with self-organized solutions strengthened by new market entries.

A wider use of credit cards and debit cards has already demonstrated that cash-based transactions are amazingly reduced. A move from physical to virtual payment tools is already providing sizable benefits to interested partners, such as customers, merchants, financial institutions, and telecom companies. There still remains uncertainty, however, in the adoption and use of mobile payment, and Ondrus and Pigneur [16] found the major reasons of such uncertainty to be lack of market maturity and the lack of policy standards.

2.2 Mobile Payment and Smart Payment

Smart payment is receiving more attention as smart phones have become popular. Therefore, this study will investigate transactions using mobile payment with a focus on smart payment, which has become a common payment method.

With the fast growth of mobile communications, smart payment has become more useful in the banking and financial industries, particularly with regard to bill payment. Such new converging technology has brought benefits to diverse partners, including banking and financial institutions and mobile communications companies and providers [11]. Also, researchers have begun to focus more on the importance of merchants to understand the functions of the mobile payment market [13, 20] as they have realized that the increasing number of these merchants has led to the success of mobile payment solutions [17].

Some say that it will be difficult for mobile payment to become the standard because of the complexity caused by the different interests arising from major partners, including financial and mobile communication businesses, which all consider mobile payment as an innovative business idea [11].

Markides [14] discusses radical product innovation rather than disruptive technologies. In his research, mobile payment was viewed as a radical product innovation, which may become disruptive for both the consumer and the producer. According to Markides [14], such innovations have been propelled by the distributors, not by the consumers.

This study assumes that smart payment is available only on smart phones, quite different from general mobile payments. Smart phones that enable convenient and secure mobile commerce services, such as electronic wallet, electronic payment, 3G broadband Internet access, and multimedia content, are then described [4]. Among these advantages, this study focuses on electronic payment functions supported by smart phones. Smart phones are capable of providing several features of smart payment, such as real-time support for segmented and personalized shopping information, credit card payment and mileage accumulation, and integration of advertising and coupons. Moreover, smart phones can provide many advantages supported by mobile payments.

2.3 Multi-Agent System

The multi-agent system is composed of multiple interacting intelligent agents. Multi-agent systems have a feature in which individual intelligent agents with diverse goals and abilities are utilized to solve problems [12, 15]. The agents are considered autonomous entities, such as software programs or robots. The interactions between them may be anti-social or cooperative, meaning that they may have a common goal or might pursue their own interests. To summarize what has been discussed so far, the features for a multi-agent system are as follows. Each individual agent has a limited capacity due to its incomplete information and ability to solve a problem, illustrating that no single agent can control the entire system. The data in this type of system is decentralized, leading to asynchronous computation capabilities.

Meanwhile, more research on the study trends related to multi-agent systems exists. Multi-agent systems have suggested a new generation of coordination within corporations [3, 21]. Also, Palmer [18] conducted a simulation test using a multi-agent system on a category-elaboration model to calculate the diversities and the accomplishments of work groups. Hahn et al. [10] studied social reputation through a flexible self-control mechanism using multi-agent system technology to determine the social simulation for the microscopic and macroscopic factors within the electronic market.

Simulation studies using multi-agent systems are actively in progress in the field of social science to coordinate organizational issues and to measure accomplishments.

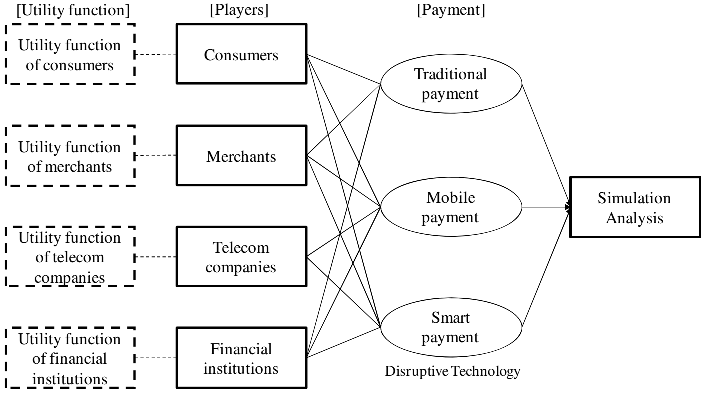


Fig. 1. Research Model

3 Research Methodology

3.1 Research Models

Our research model, as shown in Figure 1, assumes that there exist four players in a payment market, consumers, merchants, financial institutions, and telecommunication companies. Three types of payment methods are included, traditional payment, mobile payment, and smart payment. The technique adopted for this study is a multi-agent simulation (MAS).

3.2 Scenario

(1) Financial Institutions
The utility function of a financial institution (U_F) is described in Table 1.

Table 1. Utility Function of a Financial Institution

Item	Explanation
Utility function	• U_F = (TransNb x TransAvg x α TFee) + (MFee x CNb) - (MCost x CNb) - (TransNb x DiscAvg)
Detail components	• $Trans_{Nb}$ = number of transactions between consumers and merchants • $Trans_{Avg}$ = average amount per purchase (between \$0 and \$100) • α = ratio of fees between financial institutions and telecom companies (when $0.5 < \alpha \leq 1$) • TFee = merchant transaction fee (% of the purchase, e.g., 3%) • MFee = consumer credit card membership fee • C_{Nb} = number of participating consumers • M_{Cost} = maintenance cost incurred by issuing and maintaining plastic cards • $Disc_{Avg}$ = average discount amount per purchase

(2) Telecom Companies
The utility function (U_T) of a telecom company is explained in Table 2.

Table 2. Utility Function of a Telecommunication Company

Item	Explanation
Utility function	• U_T = ($Trans_{Nb}$x $Trans_{Avg}$ x (1- α)TFee) - (PT_{Cost} x M_{Nb}) - (ITD_{Cost}) + (C_{Roy} x C_{Nb})
Detail components	• $Trans_{Nb}$ = number of transactions between consumers and merchants • $Trans_{Avg}$ = average amount per purchase (between \$0 and \$100) • α = ratio of fees between financial institutions and telecom companies (when $0.5 < \alpha \leq 1$) • TFee = merchant transaction fee (% of the purchase, e.g., 3%) • PT_{Cost} = cost of the payment terminal • M_{Nb} = number of participating merchants • ITD_{Cost} = cost of the Information Technology • C_{Roy} = royalties of participating consumers • C_{Nb} = number of participating consumers

(3) Merchants
The utility function (U_M) of a merchant is described in Table 3.

Table 3. Utility Function of a Merchant

Item	Explanation
Utility function	• U_M = ($Trans_{Nb}$ x $Trans_{Avg}$ x $Profit_{Margin}$) - ($Trans_{Nb}$ x $Trans_{Avg}$ x TFee) - (($Send_{Cost}$ + CD_{Cost} + $Adver_{Cost}$) x C_{Nb}) + ($Marketing_{Eff}$ x C_{Nb})
Detail components	• $Trans_{Nb}$ = number of transactions between consumers and merchants • $Trans_{Avg}$ = average amount per purchase (between \$0 and \$100) • $Profit_{Margin}$ = the profit margin of the merchant • TFee = merchant transaction fee (% of the purchase, e.g., 3%) • $Send_{Cost}$ = cost incurred by shipping coupons and ads • CD_{Cost} = cost of the coupon discount • $Adver_{Cost}$ = cost of advertising • C_{Nb} = number of participating consumers • $Marketing_{Eff}$ = marketing effectiveness

(4) Consumers
The consumer utility function (U_C) is described in Table 4.

Table 4. Consumer Utility Function

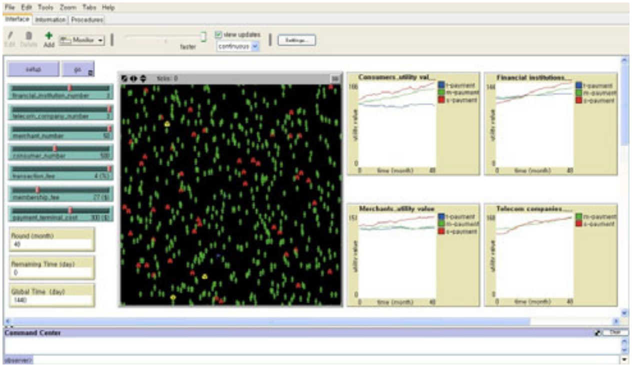
Item	Explanation
Utility function	• U_C = ($Trans_{Nb}$ x $Trans_{Avg}$ x (Pay_{Sati} + Pro_{Sati} + Dis_{Ratio} + Point)) + $Info_{Sati}$ - MFee - $Phone_{Cost}$
Detail components	• $Trans_{Nb}$ = number of transactions between consumers and merchants • $Trans_{Avg}$ = average amount per purchase (range between \$0 and \$100) • Pay_{Sati} = payment satisfaction (the degree of satisfaction with the payment function) • Pro_{Sati} = product satisfaction (the degree of product satisfaction) • Dis_{Ratio} = discount rate • Point = points (i.e., mileage) • $Info_{Sati}$ = satisfaction with shopping and card information and automatic mileage search • MFee = consumer credit card membership fee • $Phone_{Cost}$ = phone costs incurred for using mobile and smart payments

4 Experiments

4.1 Multi-Agent Simulation

Four market players are included in our simulation and are represented by multi-agents. Multi-agents are modeled using NetLogo language. Figure 2 shows a NetLogo screenshot in which each utility graph is depicted.

For the sake of performing a MAS, it is assumed that a time period is one month, and simulation time is fixed to 48 time-lag periods (i.e., 4 years). A total of 100 simulations were conducted, and the MAS results are limited to average utility values revealed by the four players during the simulation time interval. The MAS procedure adopted for this study is addressed in Figure 3.



financial_institution_number 3
telecom_company_number 3
merchant_number 50
consumer_number 500
transaction_fee 4 (%)
membership_fee 27 ($)
payment_terminal_cost 300 ($)
Round (month) 40
Remaining Time (day) 0
Global Time (day) 1440

Fig. 2. NetLogo Display

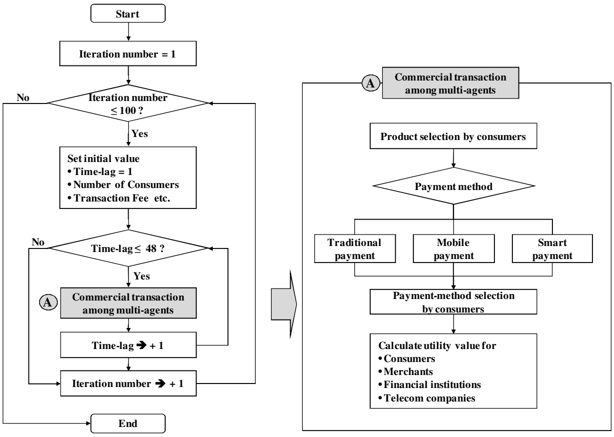


Fig. 3. Flowchart of MAS

4.2 Results and Discussion

Figure 4 shows the average utility value for the four players in three different payment methods. Figure 5 shows the trend pattern of utility value over four years (time-lag = 48). Figures 4 and 5 can be compared to induce implications.

First, at time-lag 1, consumers perceive that smart payment is the best option, followed by mobile payment and traditional payment. This fact is not surprising because smart payment provides various advantages for consumers, including personalized shopping information, card information, and an integrated payment function from which consumers can select the most convenient payment method. Moreover, through 48 time-lag periods, traditional payment shows no significant change in consumer utility value, while mobile payment does show a small change. In the case of smart payment, there is a significant increase in consumer utility value. Therefore, it can be concluded that smart payment becomes more disruptive in the market due to its significant advantages for consumers.

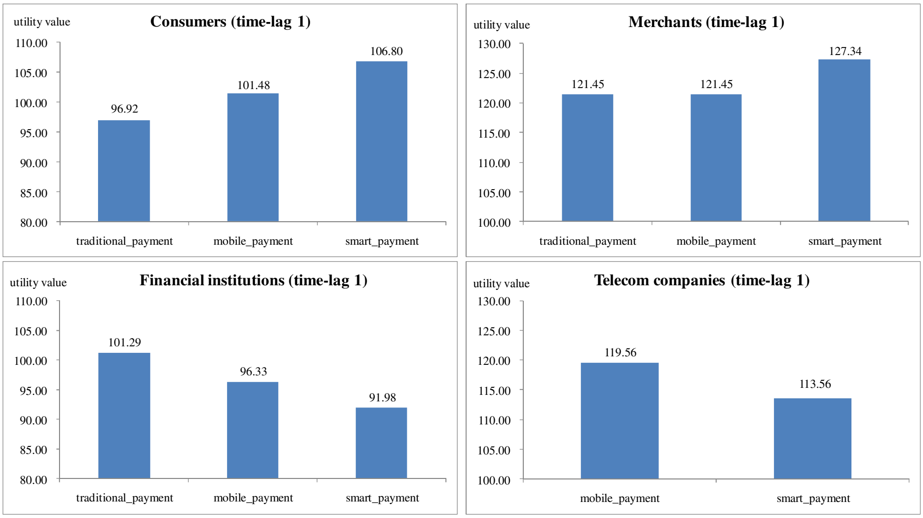


Fig. 4. Utility Value at Time-Lag 1 (Four Players and Three Payment Methods)

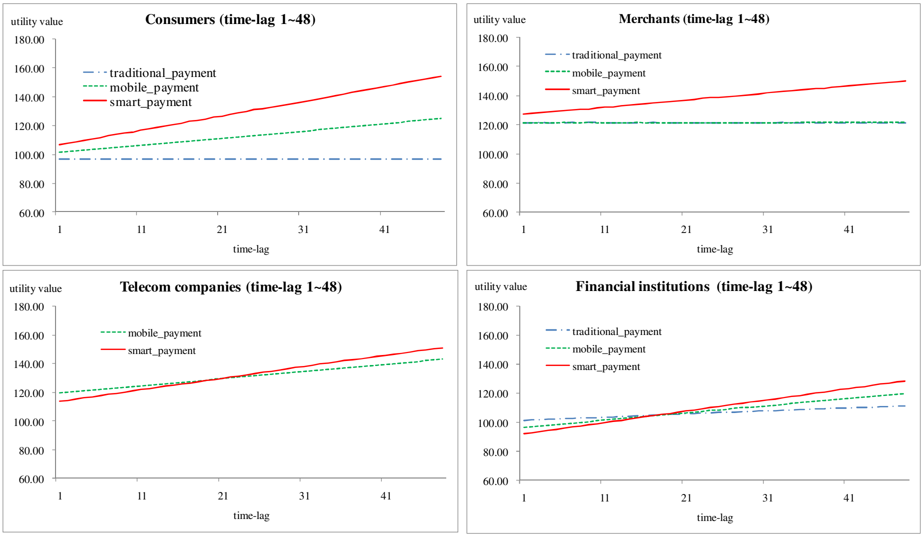


Fig. 5. Trend Patterns of Utility Value for 48 Time-Lag Periods

Second, at time-lag 1, it seems that merchants perceive the values of traditional payment and mobile payment as being nearly equivalent while those of smart payment are perceived as being more favorable because advertising costs can be saved. As with consumers, smart payment provides greater merchant utility over time.

Third, at time-lag 1, financial institutions perceive a very low value for smart payment because the transaction fee must be shared with the telecommunication companies. However, once the disadvantage of time-lag 1 is overcome, financial institutions will benefit most from smart payment due to the savings incurred by not having to issue plastic credit cards.

Fourth, at time-lag 1, telecommunication companies perceive mobile payment as the best payment method. Despite the initial IT costs that must be paid, over the 48 time-lags, smart payment will produce customer loyalty, leading to sales increases. Therefore, the long-term trend shows that smart payment is the most preferred payment method for telecommunication companies.

5 Concluding Remarks

The MAS results in this study reveal that smart payment will surely become a disruptive technology for mobile and traditional payments and that smart payment will be a very effective payment method for all four market players in the long-run. At the introductory stage, the values of smart payment were most appreciated by consumers due to its cutting-edge properties. From this study, we found that MAS can be used to derive a longitudinal pattern of effects resulting from the introduction of a new technology into the market. Further study topics include (1) integration of MAS and optimization techniques to resolve complicated decision problems and (2) the application of social network metrics among market players to analyze the economic impact of a new technology.

Acknowledgments

This research was supported by a WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant No. R31-2008-000-10062-0).

References

1. Adner, R.: When Are Technologies Disruptive? A Demand- Based View of the Emergence of Competition. Strategic Management Journal 23(8), 667–688 (2002)
2. Bower, J.L., Christensen, C.M.: Disruptive Technologies: Catching the Wave. Harvard Business Review (1995)
3. Bonarini, A., Trianni, V.: Learning Fuzzy Classifier Systems for Multi Agent Coordination. Information Sciences 136, 215–239 (2001)
4. Chang, Y.F., Chen, C.S., Zhou, H.: Smart phone for mobile commerce. Computer Standards & Interfaces 31, 740–747 (2009)
5. Charitou, C.D., Markides, C.C.: Responses to Disruptive Strategic Innovation. MIT Sloan Management Review 44(2), 55–63 (2003)
6. Christensen, C.M., Bower, J.L.: Customer Power, Strategic Investment, and the Failure of Leading Firms. Strategic Management Journal 17(3), 197–218 (1996)
7. Christensen, C.M., Raynor, M.E.: The Innovator's Solution: Creating and Sustaining Successful Growth. Harvard Business School Press, Boston (2003)
8. Christensen, C.M.: The Innovator's Dilemma. Harvard Business School Press, Boston (1997)

9. Gilbert, C.: The disruption opportunity. Sloan Management Review 44(4), 27–32 (2003)
10. Hahn, C., Fley, B., Florian, M., Spresny, D., Fischer, K.: Social Reputatiuon: a Mechanism for Flexible Self-Regulation of Multiagent Systems. Journal of Artificial Societies and Social Simulation 10(1) (2007)
11. Lim, A.S.: Inter-consortia battles in mobile payments standardisation. Electronic Commerce Research and Applications 7, 202–213 (2008)
12. Luo, X., Leung, H.: Information Sharing between Heterogeneous Uncertain Reasoning Models in a Multi-Agent Environment: a Case Study. International Journal of Approximate Reasoning 27(1), 27–59 (2001)
13. Mallat, N., Tuunainen, V.K.: Merchant adoption of mobile payment systems. In: The Fourth International Conference on Mobile Business, ICMB (2005)
14. Markides, C.: Disruptive Innovation: In Need of Better Theory. Journal of Product Innovation Management 23, 19–25 (2006)
15. McMullen, P.: An Ant Colony Optimization Approach to Addressing a JIT Sequencing Problem with Multiple Objectives. Artificial Intelligence in Engineering 15(3), 309–317 (2001)
16. Ondrus, J., Pigneur, Y.: Towards a holistic analysis of mobile payments: A multiple perspectives approach. Electronic Commerce Research and Applications 5(3), 246–257 (2006)
17. Ondrus, J., Pigneur, Y.: Cross-industry preferences for development of mobile payments in Switzerland. Electronic Markets 17(2), 142–152 (2007)
18. Palmer, V.: Simulaion of the Category-Elaboration Model of Diversity and Work-Group Performance. Journal of Artificial Societies and Social Simulation 9(3) (2006)
19. Toye, E., Sharp, R., Madhavapeddy, A., Scott, D.: Using Smart Phones to Access Site-Specific Services. IEEE Pervasive Computing 4(2), 60–66 (2005)
20. Van der Heijden, H.: Factors affecting the successful introduction of mobile payment systems. In: The 15 Bled Electronic Commerce Conference (2002)
21. Wu, D.: Software Agents for Knowledge Management: Coordination in Multi-Agent Supply Chains and Auctions. Expert Systems with Applications 20(1), 51–64 (2001)
22. Zmijewska, A.: Evaluating Wireless Technologies in Mobile Payments – A Customer Centric Approach. In: Proceedings of the International Conference on Mobile Business, ICMB 2005, IEEE Computer Society, Los Alamitos (2005)

Research on Mobile Payment Technology and Business Models in China under e-Commerce Environment*

Gu Ruijun, Yao Juan, and Wang Jiacai

School of Information Science, Nanjing Audit University, Nanjing, China, 210029
grj79@hotmail.com

Abstract. Mobile payment is one rapidly-adopting alternative payment method especially in Asia such as Japan, Korea. As 3G service gets increasingly popularized in China, Mobile payment business has evolved into its growth period. Contactless mobile payment characterized by RFID will impact e-commerce market greatly. Payment technology and business model are certain to determine the success of mobile payment. In this paper, mobile payment key enabling technology in China is analyzed and the advantages and disadvantages of various techniques are compared. We also discuss four popular mobile business models and give some typical examples implemented in China. Finally current problems and development trends in Chinese mobile payment are pointed out.

Keywords: Mobile payment, business model, RFID.

1 Introduction

The emergence of mobile commerce is affected by the current mobile networks such as 2.5G, 3G and 4G. This provides an ideal environment for payment of digital and physical goods and services. Mobile devices can be used as payment device for all types of payment situations, either electronic commerce or standard commerce. Mobile Payment (mPayment) is nowadays gaining significant attraction and many users are already using mobile devices for mobile purchase.

mPayment [1] [2] can be defined as any payment transactions involving the purchase of goods or services completed with wireless device such as a mobile phone, personal computer (wireless), or PDA. A fundamental demand for the mobile device is that it must be able to connect to a network to initiate a payment. The network could be GSM or Internet and the clearing and settlement instance could be a bank or mobile operator. The most popular concept of mPayment is users are paying from mobile phones using either prepaid or post paid methods.

mPayment is one rapidly-adopting alternative payment method especially in Asia such as Japan, Korea and China. Instead of paying with cash, cheque or credit cards, a consumer can use a mobile phone to pay for music, videos, online game subscription or items and other digital goods. As 3G service gets increasingly popularized in

[*] A shorter version of this paper has been published in a conference proceeding of FITME 2010, China.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 334–343, 2010.

China, mPayment business has evolved into its growth period. In 2009, the users of mPayment reached 108 million in China, and it is expected that the figure will be 147 million in 2010. mPayment is likely to become the first business mode integrating the Internet of Things [3] with mobile communication network.

The plan of the paper is the following: we present in Section 2 the introduction of RFID which is the enabling technology of mPayment, and we will discuss contactless mPayment technology based on RFID in Section 3. Then some popular business models of mPayment in China are compared in Section 4. At last Section 5, current problems and development trends in our mPayment are pointed out.

2 Radio-Frequency Identification

The Radio-Frequency Identification (RFID) technology [4] [5] is an automatic diagnosis technology which emerged in the 1990s. RFID is the use of an object (typically referred to as an RFID tag) applied to or incorporated into a product, animal, or person for the purpose of identification and tracking using radio waves. Some tags can be read from several meters away and beyond the line of sight of the reader. RFID comprises interrogators (also known as readers), and tags (also known as labels). Most RFID tags contain at least two parts. One is an integrated circuit for storing and processing information, modulating and demodulating a radio-frequency (RF) signal, and other specialized functions. The second is an antenna for receiving and transmitting the signal.

There are generally three types of RFID tags: active RFID tags, which contain a battery and can transmit signals autonomously, passive RFID tags, which have no battery and require an external source to provoke signal transmission, and battery, assisted passive RFID tags, which require an external source to wake up but have significant higher forward link capability providing greater range. The principle of RFID is shown in Fig. 1.

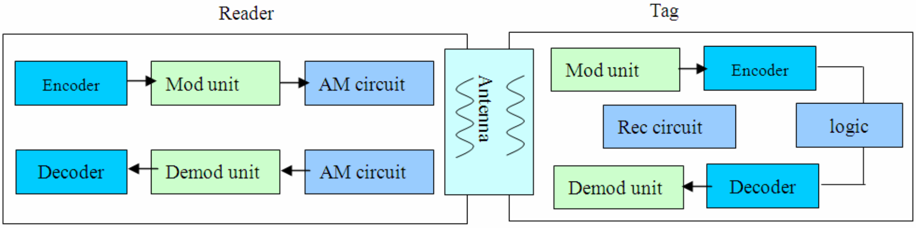


Fig. 1. Principle of RFID

As the basis of contactless mPayment application, RFID can be widely used for collecting and processing data in some areas such as logistics, transport, transportation, medical treatment, asset management, etc. In addition, RFID is the core technology of Internet of Things [3] which refers to the networked interconnection of everyday objects. It is generally viewed as a self-configuring wireless network of sensors whose purpose would be to interconnect all things. The Internet of Things will likely be a "non deterministic" and open network in which auto-organized or intelligent entities virtual objects will be interoperable and able to act independently depending on the context, circumstances or environments.

3 Contactless Mobile Payment Technology Based on RFID

Contactless mPayment technology is usually the integration of RFID cards and other cards or devices. Mobile phones integrated RFID technology mainly includes the NFC, SIMpass and RF-SIM. These three technologies provide customers with more convenience and are the support technologies for China's three major mobile operators -- China Mobile, China Telecom, China Unicom. Therefore, NFC, SIMpass and RF-SIM are the emphasis of this paper.

3.1 Near Field Communication

Near Field Communication(NFC)[6][7], founded by Phillips and supported by Nokia and Sony, is a short-range high frequency wireless communication technology which enables the exchange of data between devices over about a 10 centimetre distance. The technology is a simple extension of the ISO/IEC 14443 proximity-card standard that combines the interface of a smartcard and a reader into a single device. An NFC device can communicate with both existing ISO/IEC 14443 smartcards and readers, as well as with other NFC devices. NFC has plenty of applications, such as mobile ticketing, mPayment, Bluetooth pairing, and so on. NFC has two modes: ①Passive Communication Mode: The Initiator device provides a carrier field and the target device answers by modulating existing field. In this mode, the Target device may draw its operating power from the Initiator-provided electromagnetic field, thus making the Target device a transponder.②Active Communication Mode: Both Initiator and Target device communicate by alternately generating their own field. A device deactivates its RF field while it is waiting for data. In this mode, both devices typically need to have a power supply.

Fig.2 shows the working principle of NFC.

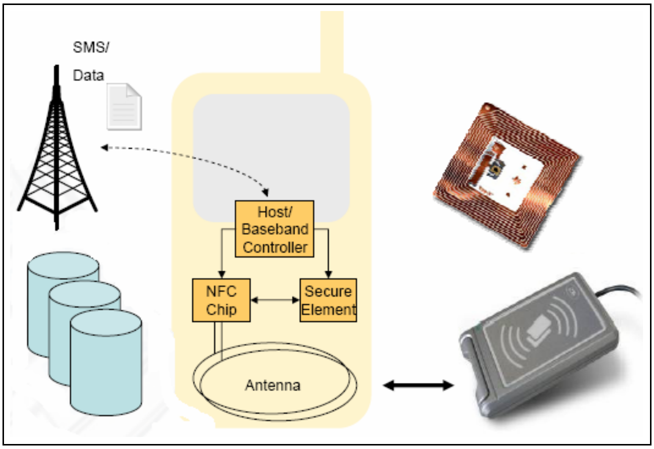


Fig. 2. Working principle of NFC

NFC technology is currently mainly aimed at being used with mobile phones. There are three main use cases for NFC: ①card emulation: the NFC device behaves like an existing contactless card. ②reader mode: the NFC device is active and read a passive RFID tag, for example for interactive advertising. ③P2P mode: two NFC devices are communicating together and exchanging information.

The NFC protocol has characteristic as follows.①secure communication and convenient establishment: The NFC protocol is one kind of short distance communication protocol, so it is safe. Just contacting both sides may establish the communication.② Supports the passive Communication model: This model is useful for the devices dependent on battery power, such as the mobile phone. ③ Compatibility: Compatible with other widely used contactless smart card protocols.

The disadvantage of using NFC mobile phone for mPayment is the high cost of hardware replacement. The user who wants to enjoy the convenience of mPayment must own a NFC mobile phone and the merchants must be equipped with appropriate payment terminals.

3.2 SIMpass

SIMpass [8] technology is founded by Watchdata who is a well-established and recognized pioneer of China in data security and smart card technology. SIMpass integrates the security module, payment module, telecom module and application module all in one SIM card. Through its contact interface, it acts as standard SIM card to execute subscriber identity authentication to your mobile phone. Via contactless interface it is ready to add contactless capabilities to your mobile phone that include transit, movie ticketing, mobile banking and payment, access control and many more applications. This allows mobile operators to stay at the centre of the mobile contactless service solution while giving their customers a highly cost effective value-added service that suit their lifestyles.

SIMpass is a Single Card Near Field Communication (SC-NFC) implementation and one of the most practical ways to implement widely accepted NFC technology. SIMpass complies with ISO/IEC 14443 Type A/B and ISO/IEC 7816 standards. The SIMpass Native card supports GSM, CDMA, China PBOC1.0, PBOC2.0 standards, Calypso European transportation standards and the Singapore CEPAS (Contactless e-Purse Application Specification) standards. Additionally, SIMpass Java card supports Java 2.2.2 and Global Platform 2.1. Both SIMpass cards also support M1 S50.

SIMpass supports telecom and non-telecom applications such as, contactless payment, e-wallet and debit and credit transactions. Compared to other NFC implementations, SIMpass has low introduction cost, as most mobile subscribers only have to change their SIM card to SIMpass card. The majority of mobile handsets currently out in the market today are compatible with SIMpass.

These are two forms of SIMpass available now: SIMpass with antenna and custom-made mobile phone. ①SIMpass with antenna: Cost effective and easily adopted by end-customers. No modifications to the mobile phones are needed. The antenna is connected to the SIMpass card and to be attached between the battery and back cover of the mobile phone.②Custom-made Mobile phones: More reliable with minimal modification needed. As an optimized SC-NFC solution, SIMpass is backed up by mobile phone vendors. There are already three models of SC-NFC mobile phones available on the market. The SIMpass antenna is either integrated into the phone battery or the main board, which can be modified by mobile phone manufacturers without making major changes on the main board.

SIMpass mainly has two drawbacks.①the user must own one mobile terminal supporting SIMpass by either replacing one new phone or transforming current phone via adding an antenna. ②the SIMpass takes up C4 and C8 ports, which usually are used for high-speed data download.

3.3 RF-SIM

RF-SIM [9] card can realize the short distant communication which is embodied with new RF technology that the users only need a smart card and make the handset they are using into an NFC-based handset with normal SIM card function.

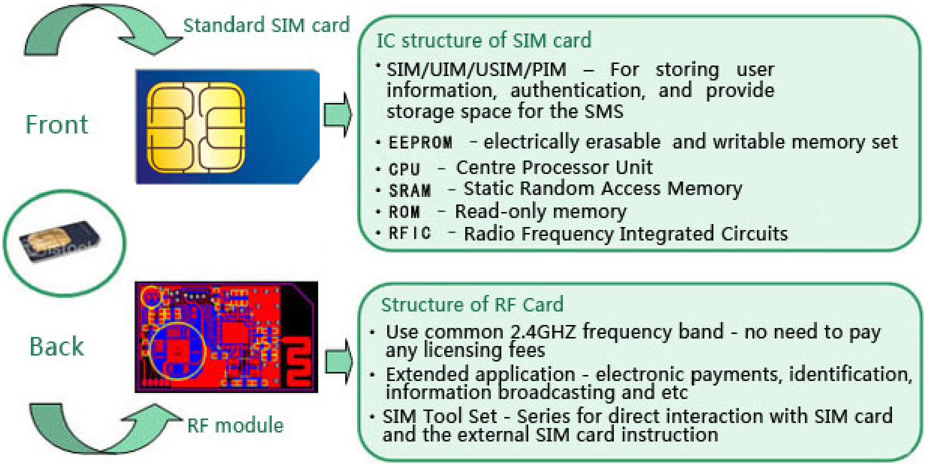

Standard SIM card
Front
Back
RF module
IC structure of SIM card
• SIM/UIM/USIM/PIM – For storing user information, authentication, and provide storage space for the SMS
• EEPROM - electrically erasable and writable memory set
• CPU - Centre Processor Unit
• SRAM - Static Random Access Memory
• ROM - Read-only memory
• RFIC - Radio Frequency Integrated Circuits
Structure of RF Card
• Use common 2.4GHZ frequency band - no need to pay any licensing fees
• Extended application - electronic payments, identification, information broadcasting and etc
• SIM Tool Set - Series for direct interaction with SIM card and the external SIM card instruction

Fig. 3. Structure of RF-SIM

RF-SIM use miniature RF modules and built-in antenna to connect the external device communication. Some SIM cards is designed for mobile phones to normal communication, authentication, and only for the physical connection; Built-in software for managing is high safety of RF-ID and other logic-based VIP membership. The structure of RF- SIM is shown in Fig.3.

Its main communication features are as follows.①Using of 2.4G frequency band, automatic frequency selection, high reliability of connection and communication. ② Two communication methods: support auto-sensing and active to connect. ③Model of two-way communication from 10CM- 500CM, can be adjusted depending on the application.④One-way data broadcasting (radius 100M).□Air transport and auto TDES data encryption, anti-eavesdropping data, the mutual authentication conducts when card accessing.

One key advantage of RF-SIM is that it can be easily retrofitted to existing mobile phones. However, since it operates at 2.4GHz rather than NFC's 13.56MHz, RF-SIM terminals are incompatible with NFC terminals.

The above three kind of technologies are now most popular in China and the main features of them are shown in Table 1.

Table 1. Comparison of three technologies (solutions)

Technology	Support mode	Security	Terminal	Frequency	Compatibility	Cost	Example
NFC	all	security	replace phone	13. 56MH	yes	high	China Telecom, China UnionPay
SIMpass	card model		transform phone			medium	China Mobile
RF-SIM		security risk	change SIM card	2.4GH	no	low	China Mobile

4 Business Models of Mobile Payment

Due to different development level and diverse industry structure among different countries, four business models [10][11] for mPayment has emerged including Operator-Centric Model, Bank-Centric Model, Collaboration Model and Third Party Model.

4.1 Operator-Centric Model

The mobile operator acts independently to deploy mPayment service. The operator could provide an independent mobile wallet from the user mobile account. A large deployment of the Operator-Centric Model is severely challenged by the lack of connection to existing payment networks. Mobile network operator should handle the interfacing with the banking network to provide advanced mPayment service in banked and under banked environment. Pilots using this model have been launched in emerging countries but they did not cover most of the mPayment service use cases. Payments were limited to remittance and airtime top up. Now, China Telecom has adopted this mode to develop mPayment though there are many difficulties needed to be solved.

4.2 Bank-Centric Model

In this model, the financial institutions take the center stage and are similar to current credit card system. A bank deploys mPayment applications or devices to customers and ensures merchants have the required point-of-sale acceptance capability. Mobile network operator are used as a simple carrier, they bring their experience to provide QoS assurance. The merchant acquiring banks and issuer banks could be different and the payment network could be managed by yet another financial institution like Visa or MasterCard. This model leverages the existing card payment system. China UnionPay is one practitioner of this model.

4.3 Collaboration Model

This model involves collaboration among banks, mobile operators and a trusted third party. Collaboration Model is seen as most feasible because it allows the stakeholders to focus on their own core competencies, opens the door for new revenue from incremental services, drives customer retention and loyalty, and responds to fundamental demand from customers. In a survey conducted by Smart Cards Alliance, 86% respondents supported this model as having the greatest potential for long term success. However, there are complicated relationships and hence complexity in negotiating deals amongst players. China Mobile China Mobile and SPD Bank have been in partnership to develop an online payment service that would be like having an Octopus card-like system in a mobile phone.

4.4 Third Party Model (Peer-to-Peer Model)

The mPayment service provider acts independently from financial institutions and mobile network operators to provide mPayment. The 3rd party company acts as a conduit between the customers, merchants and the bankers. The transaction is done

Table 2. Comparison of four business model for mPayment

Business model	Payment account	Main features	Description	Instance
Operator-Centric Model	Mobile fee account	The Operator contacts with users directly without the participation of banks, so the technology cost is low. But it is not suitable for large-value payments	Operators without payment license involved in financial transactions	China Telecom
Bank-Centric Model,	Bank account	Bank provides payment and transaction service and operator only provides information exchange. It is Suitable for large-value payments	Customized mobile phones and high initial cost	China UnionPay
Collaboration Model	Bank account	It combines Operator-Centric Model with Bank-Centric Model and is suitable for large-value payments	Operators obtain payment license indirectly through bank and is more competitive	China Mobile &SPD Bank
Third Party Model.	Bank account	The Third Party platform is a integration of merchants, banks and operators. Income is divided in accordance with their respective roles.	The third-party platform is demanding	Alibaba Pay Treasure

Peer-to-Peer between the customer and the merchant. This model is significantly different from the other three models and it threatens to eliminate the existing payment ecosystem as the role of the banks and the payment networks gets diminished. Moreover, the money can be transferred from one person to another in this way. Hence this model impacts the business of money transfer. Alibaba Pay Treasure for mobile phone is one good example of this model.

In China, there exist examples of each business model for mPayment as shown in Table 2.

5 Problems and Tendency

mPayment is typical technology-driven industry and has sufficient development space. Compared with Japan and Korea, China started later in mPayment but develop fast. At present, many problems have appeared in mobile technology and operation pattern. ①The payment security is the essential bottleneck of the mPayment. ② Different technical standards are difficult to come to an agreement. ③It is pool in mPayment service sources and is not attractive to users. ④Division of industry chain and profit distribution are urgent to be design and build.

Though there are some problems in mPayment development in China, mPayment business has evolved into its growth period [12], as 3G service gets increasingly popularized in China. mPayment is likely to become the first business mode integrating the Internet of Things with mobile communication network. Three major mobile operators together with financial institutions are actively promoting mPayment services. In May 2009, China Unicom launched its first payment service based on 3G network in Shanghai, the service was mPayment service on the basis of Near Field Communication (NFC) technology. In March 2010, China Mobile announced to acquire 20% stake of Shanghai Pudong Development Bank for the purpose of building a

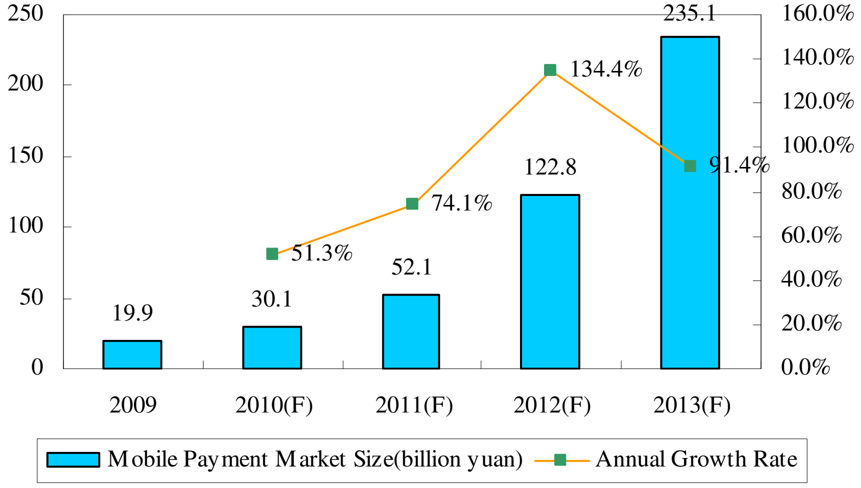


Fig. 4. Mobile market volume in China forecasting 2009-2013 [13]

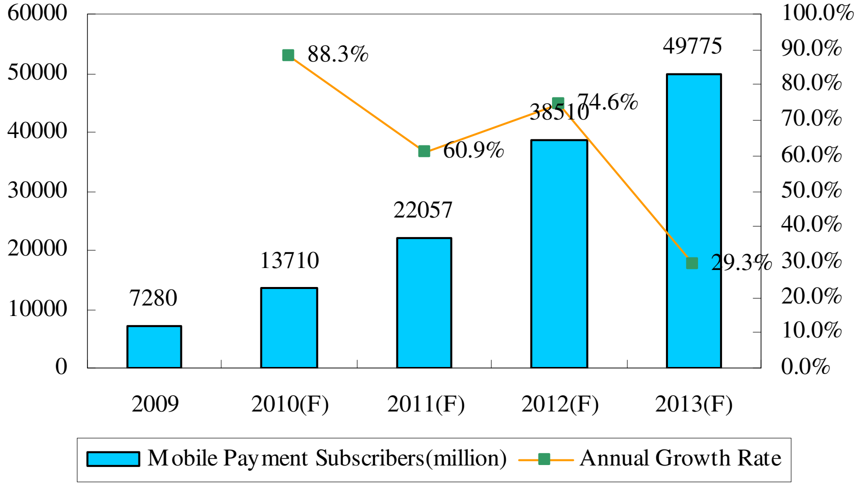


Fig. 5. Mobile payment subscribers volume in China forecasting 2009-2013 [13]

financial payment platform for developing mPayment services in the future. In addition, the huge potential market created by mPayment has attracted a large number of participants in addition to financial institutions and operators.

2010 is regarded as the beginning year of China's mPayment. Variety of innovative technologies is adopted in these trials. According to the recent research [13] as Fig.4-5 shows, the users of mPayment have reached 72.8 million at the end of 2009, and the number will be 220 million by 2011 which means triple in two years. It is predicted the market size will grow rapidly in the next few years, from 3 billion yuan in 2010 to 23.5 billion yuan in 2012. What is particularly worth mentioning is that the 3G subscribers will exceed 150 million by 2011 according to Chinese government's plan. The environment for the boost of mPayment is ripening. Mobile subscribers have exceeded 700 million in china. 3G subscribers will reach 150 million by 2011 in china. The market size will expand nearly 100% annually in average in the next three years.

6 Conclusions

In this paper, mobile payment key enabling technology in China was analyzed and compared, then four popular mobile business models were discussed. Driven by the demands of e-commerce, China's mPayment market is growing up fast. With the cultivation of user's habits and the improvement of security technology, its potential energy will be released and it also will speed up the development of huge e-commerce market of China. Mobile business, as one branch of e-commerce, is certain to develop fast with the suitable business mode reaches maturity.

Acknowledgement

This work is supported by Nanjing Audit University Research Foundation (No.NSK 2008/B10). The first author is also supported by National Funds of Social Science (No.10BGL016) and Jiangsu University Natural Science Research Foundation (No. 09KJD520006).

References

1. Karnouskos, S.: Mobile payment: a Journey through Existing Procedures and Standardization Initiatives. IEEE Communications Surveys & Tutorials 6, 44–66 (2004)
2. Gross, S., Müller, R., Lampe, M., Fleisch, E.: Requirements and Technologies for Ubiquitous Payment. In: Proc. of Techniques and Applications for Mobile Commerce, Essen, Germany (2004)
3. Yan, L., Zhang, Y., Yang, L.T.: The Internet of Things: From RFID to the Next-Generation Pervasive Networked Systems. Auerbach Publications (2008)
4. Vermesan, O., Grosso, D., Dell'Ova, F., Prior, C.: Emerging RFID Technology Roadmap. In: Proc. of EU RFID Forum 2007 Conference, Brussels, Belgium (2007)
5. Koskela, M., Ylinen, J., Loula, P.: A Framework for Integration of Radio Frequency Identification and Rich Internet Applications. In: Proc. of 29th International Conference on Information Technology Interfaces, Cavtat, Dubrovnik, Croatia (2007)
6. Timo, K., Carluccio, D., Paar, C.: An Embedded System for Practical Security Analysis of Contactless Smartcards. In: Sauveron, D., Markantonakis, K., Bilas, A., Quisquater, J.-J. (eds.) WISTP 2007. LNCS, vol. 4462, pp. 150–160. Springer, Heidelberg (2007)
7. NFC Data Exchange Format Technical Specification (2006), http://www.nfc-forum.org/specs/spec_list
8. Watchdata, http://www.watchdata.com.cn/product/html/10125.html
9. Directel Holdings, http://www.directel.cn/function.html
10. Agrawal, M.: Mobile Payments Business Models, http://www.telecomcircle.com/2009/03/mobile-payments-business-models/
11. Camponovo, G., Pigneur, Y.: Business Model Analysis Applied To Mobile Business. In: Proc. of 5th International Conference on Enterprise Information Systems, Angers, France (2003)
12. China Mobile Payment Industry Report 2009-2010. ResearchInChina (2010)
13. China Mobile Payment Comprehensive Market Report. Enfodesk (2010)

Fusion of Gaussian Mixture Densities for Face and Ear Biometrics Using Support Vector Machines

Dakshina Ranjan Kisku[1,*], Phalguni Gupta[2], Jamuna Kanta Sing[3], and Mita Nasipuri[4]

[1] Department of Computer Science and Engineering,
Asansol Engineering College
Asansol – 713305, India
[2] Department of Computer Science and Engineering,
Indian Institute of Technology Kanpur
Kanpur – 208016, India
[3, 4] Department of Computer Science and Engineering,
Jadavpur University
Kolkata – 700032, India
drkisku@ieee.org, pg@cse.iitk.ac.in,
{jksingh,mnasipuri}@cse.jdvu.ac.in

Abstract. This paper presents a multimodal biometric system for face and ear biometrics which convolves face and ear images with Gabor wavelet filters for extracting enhanced Gabor features from the corresponding images which are characterized by spatial frequency, spatial locality and orientation. Gaussian Mixture Model (GMM) is applied to the Gabor responses for measurements and Expectation Maximization algorithm is used to estimate density parameters in GMM. It produces two sets of feature sets which are fused using Support Vector Machines. Experiments on two different databases reveal its usefulness towards robust multimodal fusion.

Keywords: Multimodal biometrics, Face, Ear, Gabor wavelet filter, Gaussian Mixture Model, Support Vector Machines.

1 Introduction

Unimodal biometric systems may not be able to meet the desired performance requirements due to lack of viable characteristics. Advances in biometrics security have increased the possibility of using identification system based on multiple biometrics identifiers to combat efficiently with counter spoofing of unauthorized users. Multimodal biometric system integrates multiple sources of information obtained from different biometric cues. It takes advantage by collecting the relevant constraints together from individual biometric matchers by validating its pros and cons independently. It can overcome some of the limitations in single biometrics by fusing individual sources of information together. Experimental result reflects that the identity established by such an integrated biometric system is more reliable than that due to

[*] Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 344–351, 2010.

the single biometrics. There exist some multimodal biometrics with various levels of fusion [1], [2], namely, sensor level, feature level, matching score level, decision level and rank level. They have been found advantages over monomodal biometrics. In [3] a novel fusion approach of face and voice has been proposed where hyperbolic tangent is used for normalization and weighted geometric average is used for fusion. In [4] a multimodal biometrics fusion of face and voice with several fusion techniques has been discussed. A set of statistical learning and neural network based fusion strategies has been proposed in [5] for face and speech. A set of three score level fusion strategies for face, fingerprint and hand geometry has been presented in [6]. In [2] a fusion approach has been proposed at feature level showing significant improvements in experimental results. However, they are lacking in some respects such as robust feature extraction techniques and fusion strategies. Further, features are not well-characterized and fusion techniques are not working properly for the change in probabilities of data distributions.

This paper has proposed a fusion strategy of face and ear biometrics using Support Vector Machines (SVM). The technique uses Gabor wavelet filters [7] for convolution with the face and ear images. Gabor wavelet filters extract facial features and ear features as wavelet coefficients from the spatially enhanced face and ear images respectively where each feature point is characterized by spatial frequency, spatial location and orientation. These characterizations are viable or robust to the variations that occur due to facial expressions, pose changes and non-uniform illuminations. GMM [8] is applied to the Gabor face and Gabor ear responses for further characterization to create measurement vectors of discrete random variables. These two vectors of discrete variables are fused using SVM. Fusion of density parameters using SVM [8], [9] depends on the decision function in feature spaces. We validate the technique using two databases, each containing face and ear images. These databases are IITK multimodal database [15] and a database consisting of BANCA face dataset [13] and Technical University of Madrid (TUM) ear dataset [14]. Experimental result exhibits better accuracy obtained from the fusion approach.

The paper is organized as follows. Section 2 presents face and ear image localization and Gabor wavelets extraction. Next section discusses density estimation using GMM and Expectation-Maximization (EM) algorithm. Fusion of mixture densities using SVM is presented in Section 4. Experimental results are discussed in Section 5. Concluding remarks are made in Section 6.

2 Subject Localization and Gabor Wavelets

To locate the facial region for feature extraction, three landmarks positions on both the eyes and mouth are automatically localized by applying the technique proposed in [10]. A rectangular region is formed around the landmarks positions for Gabor characterization. This rectangular region is then cropped from the original face image which is constituted by facial part itself and background. For localization of ear region, triangular fossa and antitragus [11] are detected manually on ear image. Ear localization technique [12] has been used. Using these landmarks positions, ear region is cropped from ear image. After geometric normalization, image enhancements are performed on face and ear images. Histogram equalization is done for photometric normalization of face and ear images having uniform intensity distribution.

In the proposed approach, the evidences are obtained from the GMM [8] estimated scores which are computed from spatially enhanced Gabor face and Gabor ear responses. Two-dimensional Gabor filter [7] refers to a linear filter whose impulse response function is defined as the multiplication of harmonic function and Gaussian function. The Gaussian function is modulated by a sinusoid function. The Fourier transform of a Gabor filter's impulse response is the convolution of the Fourier transform of the harmonic function and the Fourier transform of the Gaussian function. Gabor function [7] is a non-orthogonal wavelet and can be specified by the frequency of the sinusoid and the standard deviations in x and y directions.

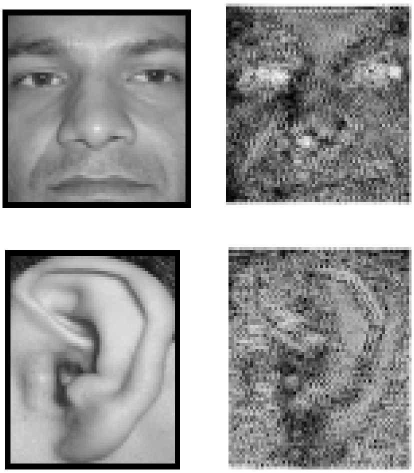


Fig. 1. Gabor Responses of Face and Ear Images

For the computation, 180 dpi gray scale images with the size of 200 × 220 pixels are used. For Gabor face and Gabor ear representations, face and ear images are convolved with the Gabor wavelets [7] for capturing substantial amount of variations among face and ear images in the spatial locations in spatially enhanced form. Gabor wavelets with five frequencies and eight orientations are used for generation of 40 spatial frequencies. Convolution generates 40 spatial frequencies in the neighbourhood regions of the current spatial pixel point. For the face and ear images of size 200 × 220 pixels, 1760000 spatial frequencies are generated. Infact, the huge dimension of Gabor responses could cause the performance degradation and slow down the matching process. In order to validate the multimodal fusion system GMM [8] further characterizes these higher dimensional feature sets of Gabor responses and density parameter estimation is performed by Expectation-Maximization (EM) algorithm [8]. For illustration, a face and an ear image from IITK multimodal database and their corresponding Gabor face and Gabor ear responses are shown in Fig. 1.

3 Density Estimation

In order to obtain better accuracy and performance, Gaussian mixture models (GMM) [8] representation has been used for the feature refinement in the proposed fusion for face and ear biometrics. The feature vectors extracted from Gabor face and Gabor ear responses can be further characterized by Gaussian distribution. Quantitive measurements for face and ear are defined by two parameters: mean and standard deviation or variability among features. Suppose, the measurement vectors are the discrete random variable x_{face} for face and the variable x_{ear} for ear. For the general case, where the feature vectors obtained from face and ear are multidimensional, the probability density function of the normal distributions is Gaussian functions [8]:

$$p(x_{face/ear}, \mu_{face/ear}, \Sigma) = \frac{1}{\sqrt{(2\pi)^{L_{face/ear}} \left|\Sigma_{face/ear}\right|}} \exp^{\frac{(x_{face/ear} - \mu_{face/ear})^T}{2\Sigma(x_{face/ear} - \mu_{face/ear})}} \tag{1}$$

where μ is the mean, Σ is the covariance matrix and L is the dimension of feature vector. Covariance matrix is the generalization to higher dimensions of the concept of the variance of a random variable. If the random variable measurements are not characterized by simple Gaussian distribution, it can be defined with multiple Gaussian components, called Gaussian Mixture Models (GMM) [8]:

$$p(x_{face}) = \sum_{m=1}^{M} \pi^{m} p(x_{face}, \mu_{face}^{(m)}, \Sigma_{face}^{(m)}) \tag{2}$$

and

$$p(x_{ear}) = \sum_{m=1}^{M} \pi^{m} p(x_{ear}, \mu_{ear}^{(m)}, \Sigma_{ear}^{(m)}) \tag{3}$$

where M is the number of Gaussian mixtures and $\pi^{(m))}$ is the weight of each of the mixture. The model of each user is the final values of $\pi^{(m)}$, $\mu^{(m)}$, $\Sigma^{(m)}$ and M, which increase the database size.

In order to estimate the density parameters of GMMs, the Expectation Maximization algorithm (EM) is adopted [8]. Each EM iteration consists of two steps – Estimation (E) and Maximization (M). The M-step maximizes a likelihood function which is refined in each iteration by the E-step.

The GMM parameters can be divided into two categories: one contains the individual mixture densities by incorporating the prior probabilities, whereas the other one contains the kernel parameter defining the form of mixture density.

4 SVM Fusion of Mixture Densities

The principle of SVM [8], [9] relies on a linear separation in a high dimensional feature space where data are mapped to consider the eventual non-linearity of the problem. To get a good level of generalization capability, the margin between the separator hyperplane and the data is maximized. A SVM classifier is trained with matching score vectors m_i, each of dimensions M. The decision surface for pattern classification is as:

$$f(\boldsymbol{m}) = \sum_{i=1}^{M} \alpha_i y_i K(\boldsymbol{m}, \boldsymbol{m}_i) + b \tag{4}$$

where α_i is the Lagrange multiplier associated with pattern $\boldsymbol{m}_i$ and $K(\cdot,\cdot)$ is a kernel function that implicitly maps the matching vectors into a suitable feature space. If $\boldsymbol{m}_k$ is linearly dependent on the other support vectors in feature space, i.e.

$$K(\boldsymbol{m}, \boldsymbol{m}_k) = \sum_{\substack{i=1\\ i\neq k}}^{M} c_i K(\boldsymbol{m}, \boldsymbol{m}_i) \tag{5}$$

where c_i are scalar constants, then the decision surface (1) can be written as

$$f(\boldsymbol{m}) = \sum_{\substack{i=1\\ i\neq k}}^{M} \alpha_i y_i K(\boldsymbol{m}, \boldsymbol{m}_i) + b \tag{7}$$

From Equation (7), one can get decision function is

$$D(f(\boldsymbol{m})) = sign\left\{ \sum_{\substack{i=1\\ i\neq k}}^{M} \alpha_i y_i K(\boldsymbol{m}, \boldsymbol{m}_i) + b^* \right\} \tag{8}$$

Equation (8) is solved for α_i and b^* in its dual form with a standard QP solver which together with decision function (4), avoids manipulating directly the elements of f and starting the design of SVM for classification from the kernel function.

In [10], the fusion strategy relies on the computation of the decision function *D*. The combined score $FS_T \in \boldsymbol{M}$ of the multimodal pattern $\boldsymbol{m}_T \in \boldsymbol{M}^R$ can be calculated as:

$$FS_T = \sum_{\substack{i=1\\ i\neq k}}^{M} \alpha_i y_i K(\boldsymbol{m}, \boldsymbol{m}_i) + b^* \tag{9}$$

Parameters can be adjusted to get various operating points. These operating points and the combined scores of the entire database are used to find the recognition rate of the proposed fusion approach.

5 Experimental Results

The proposed multimodal system has been tested on two databases viz. IITK multimodal database [15] and a database which contains face images from BANCA database [13] and ear images from TUM dataset [14]. In IITK database, there are 1200 images having 2 face and 2 ear images per individual. The face images are taken under control environment with change of ±20 degree in head pose. We have used frontal view faces only with uniform lighting and illumination condition and the near consistent facial expressions. These face images have acquired in different sessions.

The ear images are captured with high-resolution camera under control environment with uniform illumination and invariant pose. The face and ear biometrics are statistically and physiologically different from each other and independent for an individual. However, both of these physiological patterns are widely accepted and challenging in multimodal biometrics. One face and one ear image for each client are labeled as target and the remaining face and ear image are labeled as probe. For determining GMM estimated scores produced from Gabor responses, we use the entire database of face and ear images. Gaussian scores are generated from each of the two biometric modalities. Using GMMs, the density scores are produced from Gabor responses.

Beside the available IIT Kanpur database [15], we have used another database which has been built with the help of BANCA face database [13] consisting of 20×52 face images obtained from 52 subjects, each having 20 face images and of TUM ear database [14] consisting of 102 ear images taken from 17 subjects. The face images are presented with changes in pose, illumination and facial expression. On the other hand, each ear image is taken with a grayscale CCD camera and has a resolution of 384×288 pixels and 256 grayscales. Six ear instances of the left profile from each subject are taken under uniform, diffuse lighting conditions and slight changes in head position.

Six face and six ear images are considered for the creation of chimeric database for the experiment. Face images of BANCA database are taken randomly. For uniform experimental setup, face and ear images are normalized by histogram equalization. Uniform resolution and scaling are applied to all the face and ear images. A total of 6×17 images is collected separately for each face and ear modality.

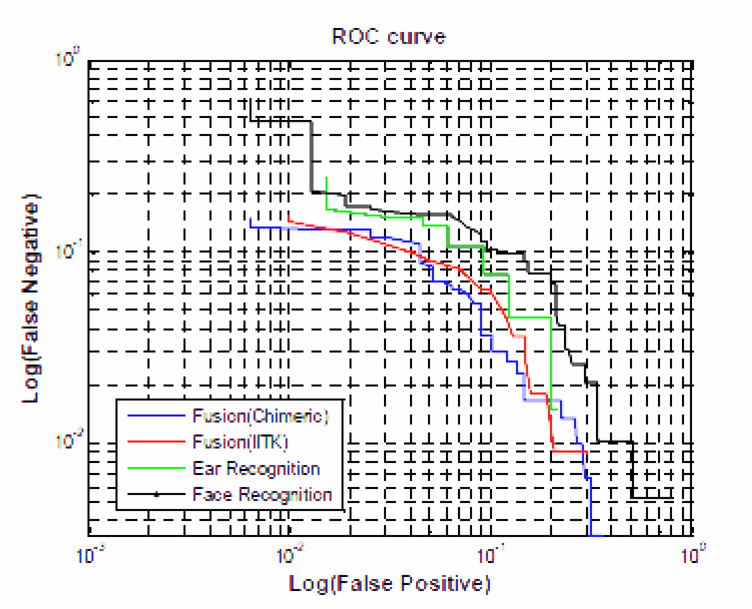


Fig. 2. Receiver Operating Characteristics Curves

Experimental results reveal that the fusion approach is better than the individual matching of face and ear biometrics. It achieves 96.49% recognition rate for IITK dataset. When other dataset is used, recognition rate is found to be 97.99%. The performance of individual face and ear matchers determined on IITK dataset only since it has viable effect to computation than that of other database. Face matcher achieves 91.96% as recognition rate. On the other hand, ear matcher achieves 93.35% recognition rate. The robust performances are exhibited when SVM based fusion uses Gabor wavelets and GMMs. ROC curves for the proposed fusion approach as well as for the individual face and ear biometrics are shown in Fig. 2. Table 1 shows Equal Error Rates (EER) and Recognition Rates (RR) for different methods which are determined on two multimodal databases.

Table 1. Equal Error Rates (EER) and Recognition Rates (RR) Determined on Two Multimodal Databases are shown

Method	Database	RR (%)	EER (%)
Multimodal Approach – I	IITK (Face + Ear Datasets)	96.49	3.51
Multimodal Approach - II	Chimeric (BANCA Face Dataset + TUM Ear Dataset)	97.99	2.01
Face Recognition	IITK Face Database	91.96	8.04
Ear Recognition	IITK Ear Database	93.35	6.65

6 Conclusion

This paper has proposed a multimodal biometrics system for face and ear biometrics. The system has been tested in two multimodal databases. Gabor filters are used to extract enhanced face and ear features which are viable and robust to different variations. E-estimator and M-estimator in GMM are used to estimate the density parameters representing the high dimensional Gabor face and Gabor ear responses. Feature sets obtained from the individual estimators are fused by SVM. Experimental results reveal its efficiency with respect to its performance for large database.

References

1. Jain, A.K., Ross, A.K.: Multibiometric Systems. Communications of the ACM 47(1), 34–40 (2004)
2. Rattani, A., Kisku, D.R., Bicego, M., Tistarelli, M.: Robust Feature-Level Multibiometric Classification. In: Proceedings of the Biometric Consortium Conference- A special issue in Biometrics, pp. 1–6 (2006)
3. Brunelli, R., Falavigna, D.: Person Identification using Multiple Cues. IEEE Transactions on Pattern Analysis and Machine Intelligence 17(10), 955–966 (1995)
4. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On Combining Classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(3), 226–239 (1998)
5. Ben-Yacoub, S., Abdeljaoued, Y., Mayoraz, E.: Fusion of Face and Speech Data for Person Identity Verification. IEEE Transactions on Neural Networks 10(5), 1065–1075 (1999)

6. Ross, A., Jain, A.K.: Information Fusion in Biometrics. Pattern Recognition Letters 24(13), 2115–2125 (2003)
7. Lee, T.S.: Image Representation using 2D Gabor Wavelets. IEEE Transaction on Pattern Analysis and Machine Intelligence 18, 959–971 (1996)
8. Bredin, H., Dehak, N., Chollet, G.: GMM-based SVM for Face Recognition. In: IEEE International Conference on Pattern Recognition, pp. 1111–1114 (2006)
9. Gutschoven, B., Verlinde, P.: Multi-modal Identity Verification using Support Vector Machines (SVM). In: Proceedings of the 3rd International Conference on Information Fusion (2000)
10. Smeraldi, F., Capdevielle, N., Bigün, J.: Facial Features Detection by Saccadic Exploration of the Gabor Decomposition and Support Vector Machines. In: 11th Scandinavian Conference on Image Analysis, pp. 39–44 (1999)
11. Iannarelli, A.: Ear Identification. In: Forensic Identification series, Paramont Publishing Company, Fremont (1989)
12. Chang, K., Bowyer, K.W., Sarkar, S.: Comparison and Combination of Ear and Face Images in Appearance- based Biometrics. IEEE Transaction on Pattern Analysis and Machine Intelligence 25(9), 1160–1165 (2003)
13. Kisku, D.R., Rattani, A., Grosso, E., Tistarelli, M.: Face Identification by SIFT-based Complete Graph Topology. In: 5th IEEE International Workshop on Automatic Identification Advanced Technologies, pp. 63–68 (2007)
14. Carreira-Perpiñán, M.A.: Compression Neural Networks for Feature Extraction: Application to Human Recognition from Ear Images. In: M.Sc. Thesis, Faculty of Informatics, Technical University of Madrid, Spain (1995)
15. Kisku, D.R., Mehrotra, H., Gupta, P., Sing, J.K.: Probabilistic Graph-based Feature Fusion and Score Fusion using SIFT Features for Face and Ear Biometrics. In: International Symposium on Optics and Photonics, vol. 7443, p. 744306 (2009)

Benchmarking Query Complexity between RDB and OWL

Chidchanok Choksuchat and Chantana Chantrapornchai

Department of Computing, Faculty of Science, Silpakorn University,
Nakhon Pathom, Thailand
cchoksuchat@hotmail.com, ctana@su.ac.th

Abstract. This paper describes how to benchmark relational database; RDB and web ontology language; OWL using query complexity concept on two difference speed machines. The domain was based on Hua Hin tourism. The purpose of this experiment was to benchmarking Semantic Web Knowledge Base Systems on relational perspective of query complexity. We use our tool to run on different speed machines to measure data complexity factors and the time of each activity. As a result, we conclude that if there is more data size and joined variables, the query complexity of RDB will increase but the ontology will reduce one. The advantage, the approach has been implemented and evaluated on improving the search engine of semantic web and reducing the expression complexity on a real archive.

Keywords: Query complexity, Ontology, OWL, RDF, RDB.

1 Introduction

The semantic web is an emerged technology. It was based on a representing, querying, and applying rules to data. The set of core standards are RDFS for structuring, RDF for representation, OWL for structuring and reasoning and SPARQL for querying. Differences from the relational databases are queried by SQL Language. The SPARQL query language [7] and protocol for RDF [10] are used as a standardized query API for providing access to datasets within enterprise settings and on the web. They aim at capturing domain knowledge and provide a commonly agreed understanding of a domain, which may be reused, shared, and run across applications and groups.

Tourism is a leading industry in the e-business. So, many projects with semantic web are already available [3], [4], [11] and [13]. The tourism enterprises will be interested in the use of ontology if they are evaluated well enough. So, the development of benchmarking ontology is an important task.

This paper contains the Hua Hin IT tourism domain covering benchmark query complexity between RDB and the knowledge base (KB) of semantic web represented by RDF/OWL. The benchmark was designed in accordance with two goals:

1. The benchmark allows comparing SQL and SPARQL query complexity across the same domain.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 352–364, 2010.

2. The benchmark is designed to measure SPARQL query performance against of OWL data.

This article makes the following hypotheses to the field of benchmarking Semantic Web technologies.

1. It complements the field with the tourism case driven benchmark.

2. It provides guidance to the developers by applying the benchmark to measure and improve the semantic web search engine by reduced the query complexity of the original RDB dataset.

The rest of the paper is structured as follows: Section 2 gives an existing benchmarks overview for Semantic Web technologies. Section 3 describes the benchmark design. Section 4 experimental presents the results of an experiment comparing the performance between RDB and OWL. Section 5 discusses the paper and outlines our next steps. Section 6 concludes the paper.

2 Literature Review

2.1 Benchmarking for RDB

Since a big part of web content is stored in RDB, the data storage becomes important. There are many types for benchmark about RDB in various domains.

In 1993, Gray [9] explained the benchmarks for DB and transaction systems. The quantitative comparison starts with the definition of a benchmark or workload that was measured as a transaction per second metric.

About the tourism domain, in 1998, a method of the performance evaluation of an information system [2] was presented in the evaluation of response time for a tourist agency's service system, result in a workload and concern to improve the system's behavior.

2.2 Benchmarking for RDF/OWL

T. Berners-Lee presented the future web concepts known as “Semantic Web” [15]. The purpose was to enable machines to comprehend semantic documents and data that are enriched by the conventional.

A key challenge for the Semantic Web is to acquire the capability to effectively query large KBs. There will be several competing systems. The benchmarks are needed that will objectively evaluate these systems. Development of effective benchmarks in an emerging domain is a challenging endeavor. For examples, LUBM: a benchmark for OWL knowledge base systems [16]. It was demonstrated with an evaluation of two memory based systems and two systems with persistent storage. It featured university domain ontology, synthetic OWL data scalable to an arbitrary size, fourteen extensional queries representing a variety of properties, and several performance metrics.

In 2007, Yuanbo et al. [8] explained to a requirements driven framework for benchmarking semantic web knowledge base systems (SWKBSs). Two major contributions were made to provide a list of requirements and organize collection of techniques and tools needed to develop such benchmarks.

In 2008, SP2Bench [12] was the language-specific SPARQL performance benchmark. It was settled in the DBLP scenario and comprised both a data generator for creating arbitrarily large DBLP-like documents and a set of designed benchmark queries. They applied existing engines as proof of concept and discussed the strengths and weaknesses from the benchmark results.

In 2009, the Berlin SPARQL Benchmark (BSBM) [5] was used for comparing the performance of native RDF stores with the performance of SPARQL-to-SQL rewriters across architectures based on e-commerce use case. The benchmark query mix emulated the search and navigation. The results of a benchmark experiment comparing the performance of four popular RDF stores (Sesame, Virtuoso, Jena TDB, and Jena SDB), two SPARQL-to-SQL rewriters (D2R Server and Virtuoso RDF Views) and two relational database management systems (MySQL and Virtuoso RDBMS).

2.3 Query Complexity Theory

Regarding to query complexity evaluation in tourism field, Abraham [1] presented the query evaluation along with business view that should not use much more of the mathematic formula and should use in tourism IT domain. The advantage of semantic web and capability of OWL reasoning in DL knowledge base is <T, A>. T is Terminological Box (TBox) constitutes the vocabulary of an application domain. A is Assertional Box (ABox) contains real world assertions of instances in vocabulary terms.

The Vardi's complexity concept [17] used ABox because ABox was viewed like RDB. It was important for the query evaluation model that was designed from KB. It meant the data model of logical query complexity itself was evaluated, rather than individual query languages. A more in-depth mathematical analysis was required for the query optimization issues that were covered by Calvanese. It takes use of DBMS techniques for both data representation, i.e., ABox assertions, and query answering via reformulation into SQL. Notably, in this case, the data complexity (ABOX size) of conjunctive query answering over ontologies is the one of First Order Logic (FOL) queries over DBs. The problematic LOGSPACE boundary was characterized. The fundamental results were presented on the data complexity of query answering in DLs. In particular, the FOL-reducibility boundary of the problem was concentrated. Query answering was no longer expressible as a FOL formula (SQL query) over the data.

There are three ways to measure the complexity of queries of evaluating queries in a specific language over a database. 1) The data complexity was given as a function of the size of the databases. 2) The expression complexity was given as a function of the length of the expressions. 3) The expression complexity was given as a function of the combined size of the expressions and the databases. It turns out that combined complexity is pretty close to expression complexity.

Vardi's definition for the complexity measure:

Definition 1: Let φ be a sentence of size s (a sentence represents a query). The φ has at most s variables. In order to evaluate φ on a database of size n, it suffices to cycle through at most n^s possible assignments of values from the database to the variables. Query complexity can be defined using formal logic as:

$$\exists\varphi\ ((\varphi \rightarrow s) \wedge (s \equiv n^{s})) \qquad (1)$$

For some sentence φ, the sentence has a size s, and s equals the number of possible values to variable assignments from the database that may be assigned to φ. Complexity of φ can therefore be expressed as the function: $s \equiv n^{s}$. The value of s (query complexity) can then be obtained by simply calculating: n^{s}.

Theorem 2: Query answering in DL is FOL reducible and therefore is in LOGSPACE with respect to data complexity [6].

3 Proposed Method

3.1 Methodology

This experiment makes the following procedure to the field of benchmarking.

1. We set up the dataset in RDB and the ontology in OWL type specified over Hua-Hin tourism. They contain the class of Attraction, Accommodation, Category, Rate, Facility, Location and Classification.

2. Set the difference levels of the complexity of conjunctive query: constant values, joining between sub-domains and variables and both of constant value and joined in the query expressions.

3. Prepare three forms of the queries: conjunctive query, SQL statement and SPARQL statement.

4. Use our tool for measuring query complexity by cover the Vardi's definition and Calvanese's theorem. We measured the query complexity both relational and knowledge-base models.

What we will investigate is as follows:

1.The results of query complexity evaluation include: number of terms, number of the first expression answers, number of the second expression answer, calculate query complexity of all expressions, degree of query complexity and percent of OWL reduced the degree of query complexity RDB.

2. The usage time of 1) on different speed machine.

3.2 Prepare the Dataset

The benchmark is settled in the Hua Hin tourism portal scenario. Figure 1 gives an overview of the dataset and the properties of each class. For RDB, the Accommodation domain was based on the structure of a normalized relational data model of http://www.huahin.go.th (2009).

After that, we created the ontology by Protégé 3.3.1 to Hua Hin tourism OWL file for SPARQL query. The DL expressivity of this ontology is SOIN (D). In OWL, classes can be described by a class identifier and can benchmark for classes [14]. The group of class and class hierarchies' benchmark contains ontologies that describe classes and class hierarchies. These ontologies include classes that are a subclass of value restrictions, cardinality restrictions on properties, and class intersections. In this group, vocabulary terms of both RDF(S) and OWL2 are used: rdfs:subClassOf, owl:Class, owl:Restriction, owl:onProperty, owl:someValuesFrom, owl:allValuesFrom, owl:cardinality, owl:maxCardinality, owl:minCardinality, owl:intersectionOf.

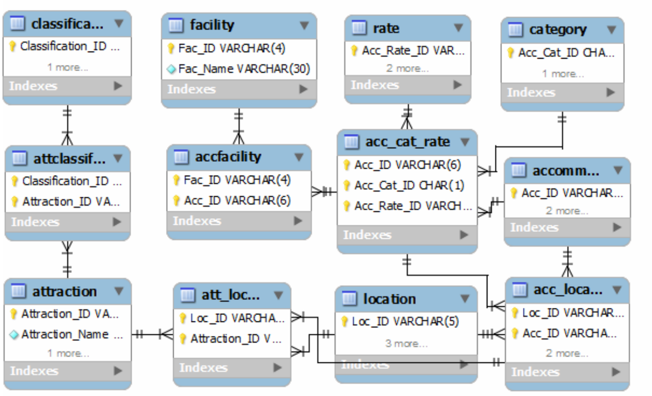


Fig. 1. The data model structure

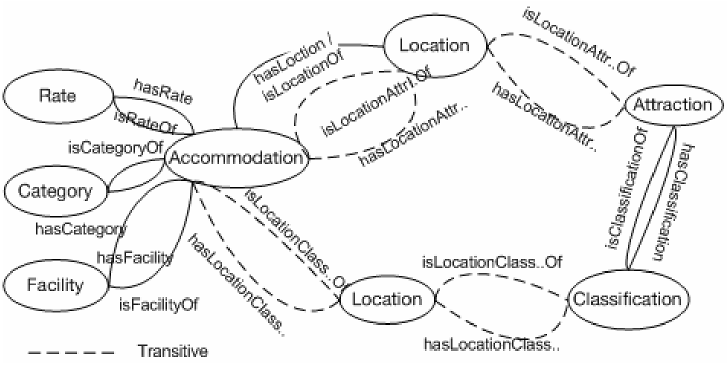


Fig. 2. OWL Ontology graph of Hua Hin tourism

In our OWL, sub-classes were use under the Accommodation class. For examples,

```
<! -- http://www.owl-ontologies.com/HuaHinProj.owl
#Category-Bangalow -->
<owl:Class rdf:about="#Category-Bangalow">
    <owl:equivalentClass><owl:Restriction>
    <owl:onProperty rdf:resource="#hasCategory"/>
    <owl:hasValue rdf:resource="#Category_Bangalow"/>
    </owl:Restriction></owl:equivalentClass>
    <rdfs:subClassOf rdf:resource="#Accommodation"/>
    <owl:disjointWith rdf:resource="#Category-
    GuestHouse"/>
    <owl:disjointWith rdf:resource="#Category-Hotel"/>
    <owl:disjointWith rdf:resource="#Category-Resort"/>
</owl:Class>
```

The overview of Hua Hin tourism ontology graph shows in Figure 2. The OWL contains the class, properties and transitive properties between classes.

3.3 Selection the Queries

The key to this research was the selection of queries that accurately reflects the structure of web queries. Since this is tourism domain from Muang Hua Hin Municipality (version 2009), the user can query the accommodation by select from category and review the destination by the location. Therefore we set up the query in select case with joining complexity. This experiment started from a basis of level-1 to level-5 that was conducted with five conjunctive queries.

Table 1. Selection of Queries

No.	Size	Basic Domain	Other Domain	Depth
1	+	/	-	0
2	++	/	Location	1
3	+++	/	Location, Attraction	2
4	+++	/	Attraction, Classification	2
5	++++	/	Location, Attraction, Classificatior	3

The queries used in the experiment are presented in Table1. It was derived by the level, data size, domain, number of joined and ontology graph depth. In the detail, we represent the conjunctive query as an ontology concept.

Test 1. Set the basic concept include the domain of Accommodation with Category, Rate and Facility. Select the Accommodation with constant value. Return the name of Accommodation from constant Rate, Category and Facility.

```
((Category ⊓  ( ∃ hasRate{ Room  Rate })
           ⊓  ( ∃ hasFacility{ Facility A })
           ⊓  ( ∃ hasFacility{ Facility B })
           ⊓  ( ∃ hasFacility{ Facility C }))
```

Test 2. Select the basic concept joined with the Location and constant values. Return the name of Accommodation from constant Rate, Category, Facility and Location.

```
((Category ⊓ ( ∃ hasRate{ Room Rate })
           ⊓ ( ∃ hasFacility{ Facility A })
           ⊓ ( ∃ hasFacility{ Facility B })
           ⊓ ( ∃ hasFacility{ Facility C })
           ⊓ ( ∃ hasLocation{ Loc_Name }))
```

Test 3. Select the basic concept joined with the Location, Attraction and constant values. Return the Accommodation name from constant Rate, Category, Facility, Location and Attraction.

```
((Category ⊓ ( ∃ hasRate { Room Rate })
           ⊓ ( ∃ hasFacility { Facility A })
           ⊓ ( ∃ hasLocation { Loc_Name })
           ⊓ ( ∃ hasAttraction { Attraction A })
           ⊓ ( ∃ hasAttraction { Attraction B }))
```

Test 4. Select the basic concept joined with the Location, Classification, and constant values. Return the Accommodation name from constant Rate, Category, Facility, Location and Classification.

```
((Category ⊓ (∃ hasLocation { Loc_Name })
           ⊓ (∃ hasRate { Room Rate })
           ⊓ (∃hasFacility  { Facility A })
           ⊓ (∃ hasFacility { Facility B })
           ⊓ (∃ hasClassification {Classification A}))
```

Test 5. Select the basic concept joined with the Location, Attraction, Classification, and constant values. Return the Accommodation name from constant Rate, Category, Facility, Location, Attraction and Classification.

```
((Category ⊓(∃ hasLocation{ Loc_Name })
           ⊓(∃ hasRate { Room Rate})
           ⊓(∃ hasFacility  { Facility A })
           ⊓(∃ hasFacility  { Facility B })
           ⊓(∃ hasAttraction {Attraction A })
           ⊓(∃ hasClassification  { Classification A })
           ⊓(∃ hasClassification  { Classification B }))
```

3.4 Measuring Query Complexity

For measuring query complexity, we used Vardi's definition as a basis for the complexity measure. Then we used Calvanese's theorem to concern about the cycle times of the actual degree of a query's computational complexity in LOGSPACE.

3.5 The Machine Specification

The experiments ran on two different speed machines as follows:

Machine#1: The experiment was conducted on a processor: Intel (R) CPU T2050 1.60 GHz; 798 MHz; memory: 0.99 GB hard disks: 80GB 32-bit Operating System running Window XP Professional.

Machine#2: The experiment was conducted on a processor: Intel (R) Core(TM) i5 CPU M430 2.27GHz; memory: 4GB hard disks: 320GB 64-bit Operating System running Window 7 Home Premium.

Software were installed in 2 machines: Apache Web Server 2.2.8, MySQL Database 5.0.51b, phpMyAdmin Database Manager 2.10.3. Protégé 3.3.1.RacerPro 1.9.0 reasoner. Java 1.6.0_18 was used.

4 Experiment Results

The experiment result was represented in Table2. We measure the factors include:

1. # Terms: Number of terms from conjunctive query.
2. # Joined: Number of equi joined between the tables.
3. Xs: the first parameter from Vardi's concept.
4. Vs: the second parameter from Vardi's concept.
5. Query complexity: Xs* Vs.
6. # Answer: the number of the answers of query.
7. Complexity Ranking: the depth of OWL graph.
8. Degree of query complexity in LOGSPACE.
9. Percent of reduced the degree of query complexity.
10. The ratio of the #answers and query complexity.
11. Usage time in millisecond on machine#1.
12. Usage time in millisecond on machine#2.
    (Usage time = load time + execution time)

Table 2. Results of benchmarking query complexity

Measuring	RDB1	OWL1	RDB2	OWL2	RDB3	OWL3	RDB4	OWL4	RDB5	OWL5
1) #Terms	7	6	8	7	9	7	9	7	11	9
2) #Joined	0	0	1	0	2	0	2	0	3	0
3) X^s	2	2	2	2	4	2	5	2	6	5
4) V^s	-	-	4	-	5	-	222	-	1110	-
5) QC	2	2	8	2	20	2	1110	2	6660	5
6) #Answer	2	2	2	2	2	2	2	2	5	5
7) Ranking	-	-	1	1	1	2	1	2	1	3
8) Deg QC	0.30	0.30	0.90	0.30	1.3	0.30	3.0	0.30	3.8	0.70
9)%Reduce of QC	0		67		77		90		82	
10) #Ans/QC	1.00	1.00	0.25	1.00	0.10	1.00	0.00180	1.00	0.00075	1.00
11) M#1(ms.)	23	4.3	24	5.4	24	5.7	26	6.8	48	8.8
12) M#2(ms.)	9.5	4.8	11	4.9	12	5.7	19	5.8	32	6.6

5 Result Analysis

We analyzed the result by the issues as follows:

Issue1: About usage time of RDB and OWL. The machine#2 can reduce usage time more than the machine#1. And the usage time of OWL less than the usage time of RDB in both machines. The differences of the usage time of OWL in the both machines less than the usage time of RDB. That means whether speed of machines, OWL still use less time than RDB clearly.

Issue2: The number of terms and joins between RDB and OWL are related as Figure 4. The RDB and OWL expression terms increase along with the more depth of graph. But the amount of OWL expression terms still increase less than RDB terms because sub-class in OWL can reduce the expression terms. We used the subclass of Accommodation class in OWL.

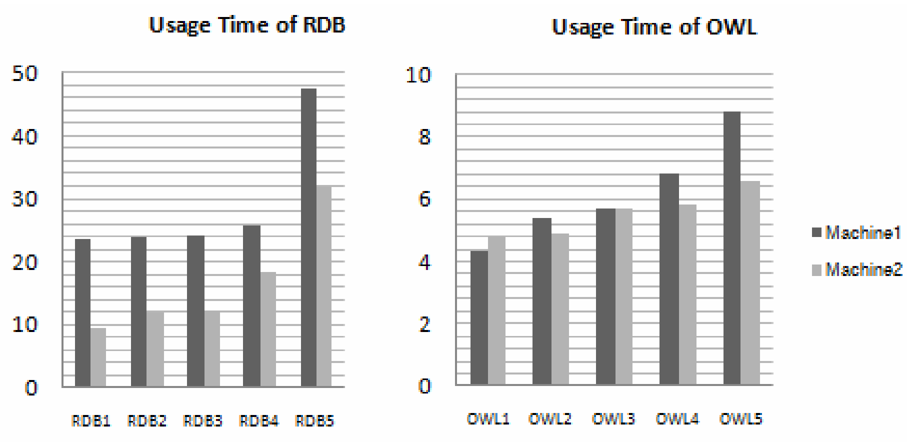


Fig. 3. Comparison between usage times of RDB and OWL

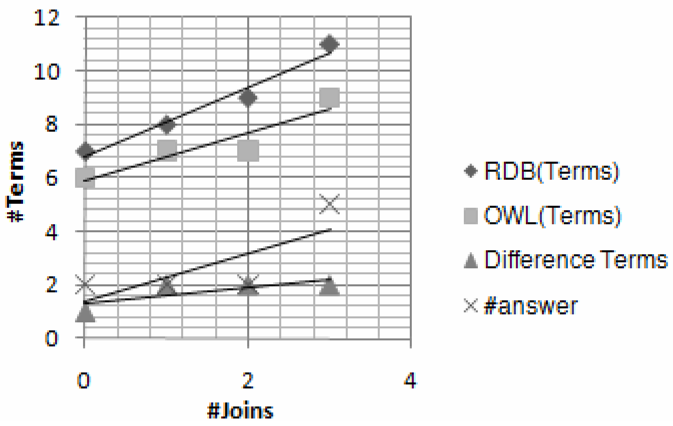


Fig. 4. Number of terms and joins of RDB and OWL

When the SPARQL were queried, it can be found out within the sub-class constantly. We can imply the OWL expression by subclass terms instead the terms of class. We can show you clearly by the conjunctive query. For example, Conjunctive query of OWL-2:

```
Q(X)<-Category-Bangalow(X) ∧ hasLocation(X,A) ∧
hasRate(X,B) ∧ hasFacility(X,C) ∧
hasAccommodationFacility(X,D) ∧
hasAccommodationFacility(X,E) ∧A=HuaHin-Takiab Road ∧
B= Room_Rate_2 ∧ C =Beach ∧ D= Refrigerator ∧ E= Air
Conditioning
```

Whereas the conjunctive query of RDB-2 have to join between Accommodation and Category classes as:

```
Q(X)<- Accommodation(X) ∧ hasLocation(X,A) ∧
hasRate(X,B) ∧ hasCategory(X,C) ∧
hasAccommodationFacility(X,D)  ∧
hasAccommodationFacility(X,E) ∧ hasFacility(X,F) ∧
A=HuaHin-Takiab Road ∧ B= Room_Rate_2 ∧ C=
Category_Bangalow ∧ D=Beach ∧ E= Refrigerator ∧ F= Air
Conditioning
```

This is the reason why the number of terms of OWL less than RDB terms in the Table 2.

In addition, by study the graph in Figure4, we can see the difference between the terms of RDB and OWL slight wider, when the number of terms and joins increase. Because transitive properties were add in OWL. For instance, the use of transitive property hasLocationAttraction in the Accommodation ontology eliminated the number of joins for a relational model, and also reduced the number of query terms. In the next line, you can see the Accommodation class connects Location class with hasLocation property. Location class connects to Attraction class with hasAttraction Property.

Accommodation →Location →Attraction

We can formulate easier if we use transitive property hasLocationAttraction to connect between Accommodation and Attraction as below:

Accommodation →Attraction

For the other transitive properties, we show in Figure 2. As a result, OWL3, OWL4 and OWL5 can reduce the expression terms from relational model. In addition, the transitive property use of reduce the query complexity as well.

Issue3: the Comparison of Query Complexity of RDB and OWL. By studying Figure 5, we can see the difference query complexity values between RDB and ontology with the same data size. The RDB represents by SQL query language, find out the answer from the joins among the table, data complexity will increase the query complexity follow by the data size. On the contrary, the OWL represents by SPARQL query language can query the answer along the transitive property and subclass without joins. The OWL query complexity does not increase that much.

Issue4: The percentage of OWL implied the reduction in the query complexity. Figure 6 shows the reduction in percent of OWL query complexity degree to that of RDB.

We can see the growth of OWL reduced the degree of query complexity from RDB when the query complexity increases. In the small data complexity case such as query 1, we cannot see the difference. But for more query complexity, the percent of OWL reduced the query complexity degree from RDB more than 80%.

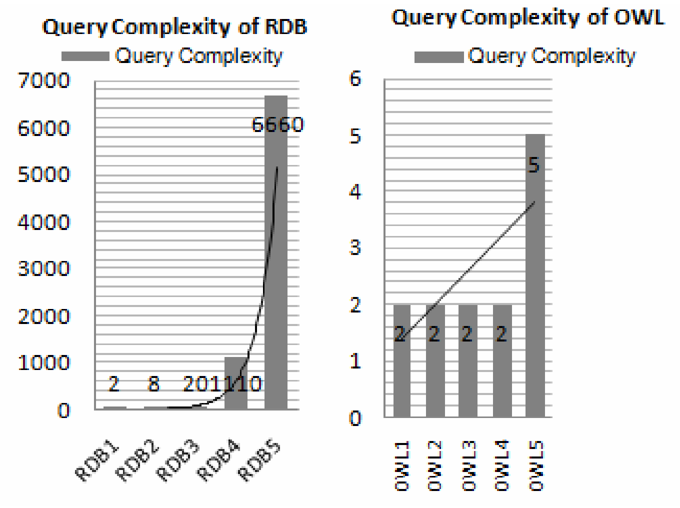


Fig. 5. Comparison of Query Complexity of RDB and OWL

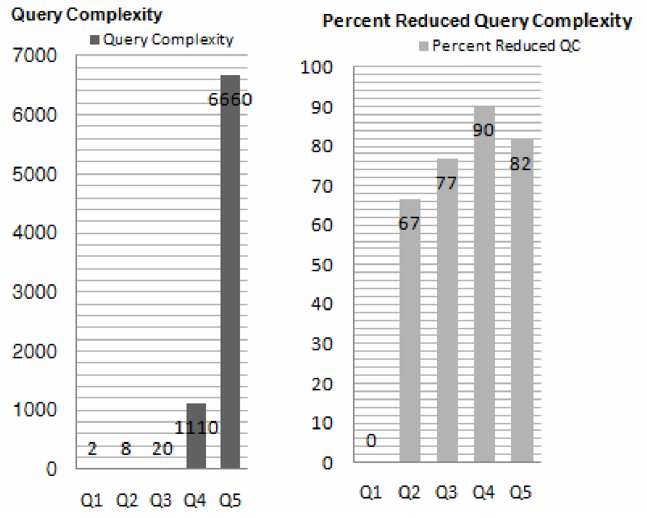


Fig. 6. Percent of reduction of OWL on the query complexity from RDB

Issue5: The ratio of number of the query answer to query complexity. We can see clearly in RDB that the query used to find the answer was much more complex. Because the number of answers were equal to 5 and query complexity as 6600. The ratio was less than 0.00076. This value was compared with the ratio equal to 1 from OWL side. When there is a small number of the answer, it is easier to find.

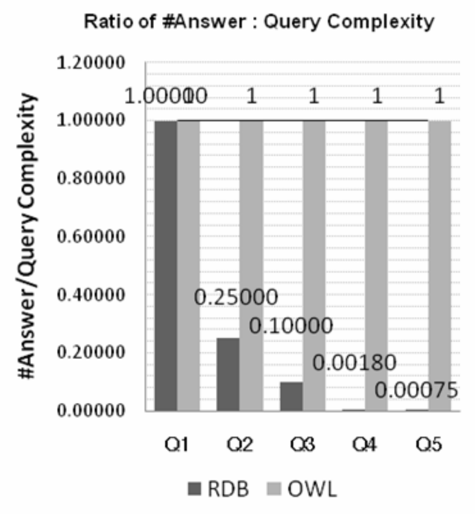


Fig. 7. The ratio of number of the answers to query complexity

6 Conclusion

Overall, we proposed the method of benchmarking query complexity between RDB and OWL. That can separate to the overview; measuring the usage time by two different speed machines. It was clear that the faster 4GBRAM machine#2 could run RDB and OWL in less usage time than the other one. The gap of time in RDB fell clearly. But in the OWL case, the Machine#2 dropped a few usage times for the small data. For the large amount of data, it tended to decrease clearly.

We also evaluated query complexity in five level queries. The RDB represented by SQL language that could query by joins between the tables. As the result, the more number of joins, the more query complexity. Whereas OWL ontology represented by SPARQL language. There were the sub-class hierarchy and transitive properties which used to reduce the query complexity. Thus, in the future work, we can use them continuously to improve the search engine in Semantic Web technologies.

Acknowledgments. This experiment is the part of thesis: Improving search engine using semantic web case study Hua-Hin tourism information. Thank you to the Muang Hua Hin Municipality for giving the http://www.huahin.go.th (2009) database structure.

References

1. Abrahams, B.: Tourism Information Systems Integration and Utilization within the Semantic Web (2006), `http://wallaby.vu.edu.au/adt-VVUT/uploads/approved/adt-VVUT20070514.125504/public/02Chapter1-3.pdf`

2. Zgrzywa, A.: The evaluation of the response time for a tourist agency's service system. Information and Software Technology 40, 37–44 (1998)
3. Legrand, B.: Semantic Web Methodologies and Tools for Intra-European Sustainable Tourism. JITT (2004)
4. Robert, B., Christina, F., Birgit II, P., Christoph, G., Hannes, W.: Covering the semantic space of tourism: an approach based on modularized ontologies. In: Proceedings of the 1st Workshop on Context, Information and Ontologies, ACM, Heraklion (2009)
5. Bizer, C., Schultz, A.: The Berlin SPARQL Benchmark. International Journal on Semantic Web and Information Systems - Special Issue on Scalability and Performance of Semantic Web Systems 5, 1–24 (2009)
6. Calvanese, D., Lembo, D., Lenzerini, M., Rosati, R.: Data complexity of query answering in description logics. In: Proc. of KR 2006, pp. 260–270 (2006)
7. Prud'hommeaux, E., Seaborne, A.: SPARQL Query Language for RDF.W3C Recommendation (2008), http://www.w3.org/TR/rdf-sparql-query/
8. Yuanbo, G., Abir, Q., Zhengxiang, P., Jeff, H.: A Requirements Driven Framework for Benchmarking Semantic Web Knowledge Base Systems. IEEE Educational Activities Department, vol. 19, pp. 297–309 (2007)
9. Gray, J.: Database and Transaction Processing Performance Handbook. In: The Benchmark Handbook for Database and Transaction Systems. Morgan Kaufmann, San Francisco (1993)
10. Clark, K.G., Feigenbaum, L., Torres, E.: SPARQL Protocol for RDF.W3C Recommendation (2008), http://www.w3.org/TR/rdf-sparql-protocol/
11. Dell'Erba, M., Fodor, O., Höpken, W., Werthner, H.: Exploiting Semantic Web Technologies for Harmonizing E-Markets. Journal of Information Technology and Tourism 7, 201–219 (2005)
12. Schmidt, M., Hornung, T., Lausen, G., Pinkel, C.: SP2Bench: A SPARQL Performance Benchmark. In: IEEE 25th International Conference on Data Engineering, ICDE 2009, pp. 222–233 (2009)
13. Foder, O., Werther, H.: Harmonise: A Step Toward an Interoperable E-Tourism Marketplace. International Journal of Electronic Commerce 9, 11–39 (2005)
14. Castro, R.G.: Benchmarking Semantic Web technology. Facultad de Informática, Doctoral Thesis. Universidad Politécnica de Madrid (2008)
15. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web: Scientific American. Scientific American (2001)
16. 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, 158–182 (2005)
17. Moshe, Y.V.: The complexity of relational query languages (Extended Abstract). In: Proceedings of the fourteenth annual ACM symposium on Theory of computing. ACM, San Francisco (1982)

FEDTIC: A Security Design for Embedded Systems with Insecure External Memory

Mei Hong and Hui Guo

School of Computer Science and Engineering,
The University of New South Wales,
Sydney, NSW 2052, Australia
meihong@cse.unsw.edu.au, huig@cse.unsw.edu.au

Abstract. This paper presents a security design for embedded systems that have a secure on-chip computing environment and an insecure off-chip memory. The design protects the confidentiality and integrity of data at a low cost on performance and memory consumption. We implemented the design based on the SimpleScalar simulation software. Our simulation on a set of benchmarks shows that very little overhead is incurred for on-chip memory, and the average overheads on performance and off-chip memory, are only 7.6% and 6.25%, respectively.

1 Introduction

Security becomes increasingly important in embedded systems. One driving force is the security-aware services that embedded systems provide, such as financial transactions, application downloading on mobile devices. Embedded software systems are vulnerable, as they are written in insecure lower-level language with poor support for runtime error checking [1]. The embedded systems can also easily be attacked through physical accesses. Therefore, the confidentiality and integrity of the data that are processed and communicated must be ensured.

Processors and memory are two typical components in the embedded system for computing and data storage. With the fast expanding of embedded system applications, the requirement for memory continues to grow. Apart from the memory equipped on the processor chip, large off-chip memory is usually indispensable. However, the link between the off-chip memory and on-chip processor is often a weak security point and the off-chip memory is more vulnerable to variety of attacks.

The conventional way for confidentiality protection is to use encryption techniques. The data exposed to possible attacks are encrypted; Only in a secure environment, can the plain data be viewed and used. When the data are stored in or transmitted to outside of the secure environment, they are always in an encrypted format.

For the integrity protection, a typical solution is tagging the data – attaching a piece of data item (or called **tag**) that is exclusive to the data to be protected. Any alteration to the data by an intruder will result in a corrupted tag, which forewarns the data should not be used.

However, realizing these security protections is often at the cost of performance, hardware area, and energy consumption. In this paper, we propose an efficient, low

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 365–375, 2010.

cost security design – a Fused Encryption/Decryption and Tagging/Integrity Checking (**FEDTIC**) engine – for systems that have a secure on-chip computing environment and insecure off-chip memory. We reduce the area overhead by sharing one hardware component for both encryption and decryption. We improve performance by combining the encryption/decryption operation and tagging/integrity checking into one step, and this step is performed in parallel with the memory access so that the impact of the security implementation on the system performance is minimized.

The paper is organized as follows. Section 2 reviews the related work. The *FEDTIC* design is explained in Section 3. The experimental setup and simulation results are presented in Section 4, and the conclusive remarks are given in Section 5.

2 Related Work

There are many types of possible attacks. An overview of them can be found in [2].

The "secure processor" with an insecure external memory was first proposed by Best [3]. According to Best, only the processor chip is secure, other external components are vulnerable to attacks. To increase the security, contents on external memory are encrypted and will be decrypted when they are fetched to the processor. The cipher unit and secret key used for encryption and decryption are kept on-chip. A set of commercial secure processors, Dallas Semiconductor DS5000 series [4], are designed based on this idea and have been used for different applications, such as the pay-TV access controller and credit card terminal.

Blum, etc. [5] proposed an approach for the integrity protection of memory data. Significant efforts have been made to enhance the security for both the confidentiality and integrity of the external memory. Confidentiality is achieved through instruction and data encryption. Integrity verification is accomplished by creating an authentication tag for each memory block using MAC (Message Authentication Code) function or cryptographic hash function. The encryption and integrity verification can be implemented in software or hardware.

Since software-only methods, such as code obfuscation and watermarking, do not resistant physical attacks, intensive researches are devoted on architectural support for secure environment. Two most-referenced approaches are XOM [6,7], and AEGIS [8,9,10]. Both adopt the Best's architecture: secure processor chip and insecure external memory.

XOM provides an architectural support for copy and tamper resistant software, where the execute-only memory(XOM) is implemented that allows instructions stored in external memory to be executed but not otherwise manipulated. Instructions and data are stored in an encrypted form. Integrity is achieved by isolating independent software applications running on the same processor. Each application is stored in a compartment for the secure execution; accessing contents in one compartment by applications from others is forbidden. Tags are created to identify contents in each application and different session keys are used to encrypt associated data.

AEGIS provides both a tamper-resistant environment where attackers are unable to obtain any information from system operation (confidentiality), and a tamper-evident, authenticated environment in which any physical or software tempering is guaranteed to be detected (integrity).

One of the problems in these architectures is the performance overhead. Every external memory transaction including both instructions and data undergoes encryption and decryption. They cannot be used until they are fetched from the external memory and decrypted. Cryptography is a computational intense operation, which greatly degrades performance. Moreover, adding tags for integrity verification results in more memory consumed.

Authors in [11] proposed a "CryptoPage" architecture which implements memory encryption, memory integrity checking and information leakage protection with a low performance penalty. The authors combined the AES encryption in the counter mode of operation with Merkle tree authentication for these encryption, integrity protection security features. The Merkle tree technique decreases the on-chip memory overhead, which is incurred by storing hash values or nonces. Elbaz et al. [12] improved the tree structure by parallelizing the hash tree update operation.

The above methods combine the fast hardware implementation for encryption and Merkle tree for integrity. This basic composition [13] generally uses Encrypt-then-Mac, a two pass approach. Therefore, the cost to achieve confidentiality-and-integrity is the cost of encryption plus the cost of MAC computation. Several one-pass approaches, such as AE modes [14] have been proposed. However, to the best of our knowledge, no implementation of such a design has been reported for external memory protection in embedded systems.

In this paper, we develop a security engine for external memory encryption and integrity verification. This engine can provide the same security level as other existing one-pass designs, but has significant low overheads on both performance and memory consumption.

3 Design Approach

Figure 1 shows the top-level architectural setting of the Fused Encryption/Decryption and Tagging/Integrity Checking (*FEDTIC*) engine in a system with an on-chip processor and an off-chip memory.

The *FEDTIC* is implemented on the processor chip (secure area). For confidentiality, data stored in the off-chip memory are all in an encrypted form (as marked inside the " [] " when they are transferred over the memory bus outside of the processor chip). The tag of data is affixed at the end in the cache line for each transfer between the on-chip cache and off-chip memory. The *FEDTIC* engine provides encryption/decryption, tagging and integrity verification for each of the external memory accesses.

We introduce a value, called Line-Access Stamp (**LAS**) that serves as a hallmark for each cache-line memory access. For every memory write, a new LAS is generated and used to encrypt and tag the cache line that is written to the memory. When read from memory, the memory data are decrypted and the related tag is calculated. The calculated tag is then compared with the tag coming along from the memory. If they are same, the integrity of the memory data is verified and the data is loaded to the cache; otherwise, the memory read operation failed.

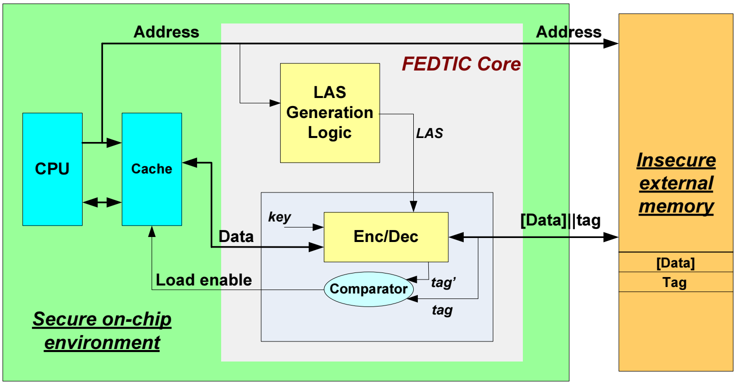


Fig. 1. System Architecture

Different from other designs, our *FEDTIC* engine features following specialities:

- The encryption and decryption operations are identical. Therefore, we use a same hardware component for both encryption and decryption, hence saving the hardware cost.
- We exploit the encryption/decryption process and let the integrity tag generation/checking mix with the encryption/decryption to speed up *FEDTIC* operation; and
- The *FEDTIC* operation is in parallel with the memory access, hence its impact on the overall system performance is reduced to a lowest possible level.

The key parts of the $FEDTIC$ design are detailed in the following subsections.

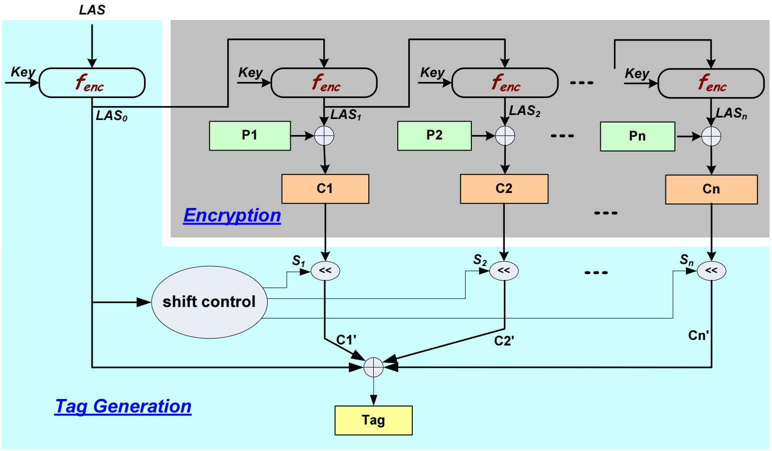


Fig. 2. Encryption and Tag Generation

3.1 Encryption and Tagging

Since the symmetric encryption can run as much as 1000 times faster than the asymmetric encryption, we employ the symmetric encryption in our design. With our design, the encryption and decryption are always performed on the processor chip, hence no need for the encryption key distribution (which would otherwise be an issue, commonly faced by many symmetric encryption applications). We use the symmetric cipher on blocks with fixed length (e.g. 64 or 128 bits). A cache line (data transferred between the cache and external memory for one memory access) consists of multiple such blocks. Therefore, we use block cipher to encrypt a cache line in the mode operation (with the Output FeedBack mode (**OFB**) as showed in Figure 2).

With the OFB mode, instead of encrypting plaintext blocks directly, LAS is recursively encrypted with a symmetrical encryption function, f_{enc}. The output of the previous encryption is the input of the current encryption. Each round of encryption produces a separate LAS value, which is XORed with the plaintext block (denoted as $P1, P2, \cdots, Pn$, respectively) to generate a ciphertext ($C1, C2, \cdots, Cn$, correspondingly), as shown in the region marked as *Encryption* in Figure 2.

Algorithm 1. Encryption process and tag generation

```
/* recursive $LAS$ encryption */
$LAS_0$ = $f_{enc}(LAS)$;
for $i = 1$ to $n$ do
    $LAS_i$ = $f_{enc}(LAS_{i-1})$;
end for
/* plaintext block encryption */
for $i = 1$ to $n$ do
    $C_i = LAS_i \oplus P_i$;
end for
/* encrypted data: concatenation of all $C_i$ */
C = Null;
for $i = 1$ to $n$ do
    $C = C \parallel C_i$;
end for
/* shift control operation for each block based on the bit segments of $LAS_0$ */
for $i = 1$ to $n$ do
    m_bit_segment = bit $m * (i - 1)$ to bit $m * i - 1$ of $LAS_0$;
    $S_i$ = number_of_zeros in (m_bit_segment);
end for
/* transformed ciphertext blocks: left shift $S_i$ bits for block $i$ */
for $i = 1$ to $n$ do
    $C'_i = C_i << S_i$;
end for
/* Tag: XOR of transformed blocks and $LAS_0$ */
tag = $LAS_0$;
for $i = 1$ to $n$ do
    $tag = tag \oplus C'_i$;
end for
```

To obtain a cache line tag, the cheap way (often used in traditional designs) is XORing all data blocks in the cache line. The data blocks can be in either a plaintext format or a ciphertext format. But either way invites potential attacks. If the plaintext was used, the original data could likely be deduced due to the easy availability of the tag provided by the insecure memory; On the other hand, if the ciphertext was used, altering data by switching different blocks in the off-chip memory would not change the tag value, hence an integrity attack would easily be applied.

We use a *transformed ciphertext blocks* (by bit shifting operations) in the tag generation. Before the XOR operation, each block is left-shifted and the number of bits to be shifted is controlled by the *Shift Control Logic* which is, in turn, determined by the encrypted *LAS* value, LAS_0, as illustrated in the second region in Figure 2. With this tag design, any swapping of the ciphertext blocks in the external memory will have a different tag value, hence the attack can be identified.

The operations in the encryption and tagging process are summarized in Algorithm 1, where we assume there are n blocks in a cache line and each block has m bits.

From Figure 2, we can see that the encryption process and tag generation can be partially parallelled. After the first round of *LAS* encryption, the shift control logic can perform in parallel with the encryption. All S_i values can be generated once LAS_0 is available; C_i can be calculated when LAS_i is completed; C_i can be transformed (for the tag calculation) as soon as S_i is known. The extra time taken by the tag generation is just the sum of the execution times for one shift and one XOR operation. Therefore, the total execution time is reduced. An parallel execution example is given in Figure 3, where a cache line of 4 blocks is assumed.

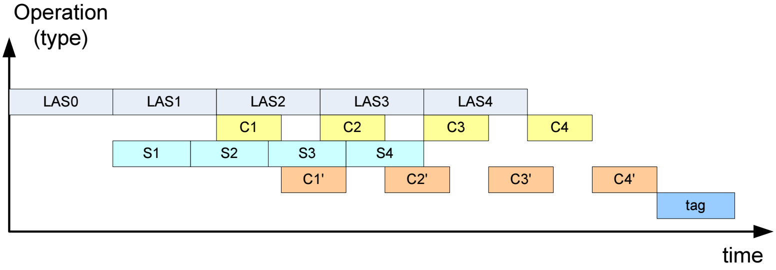


Fig. 3. Parallel Operations in Encryption and Tag Generation During Memory Write

3.2 Decryption and Integrity Checking

The decryption and integrity checking for a cache line that is fetched from memory is given in Figure 4.

It consists of two parts: decryption and tag calculation that is based on the ciphertext blocks from the memory. Because the *OFB* mode of operation is used, as for the same cache line, the same *LAS* is encrypted. The decryption is identical to the encryption process. Because of the symmetrical attribute of the XOR operation, the results of the encryption are XORed with ciphertext to get the plaintext. The tag is calculated in the same way as in the tag generation.

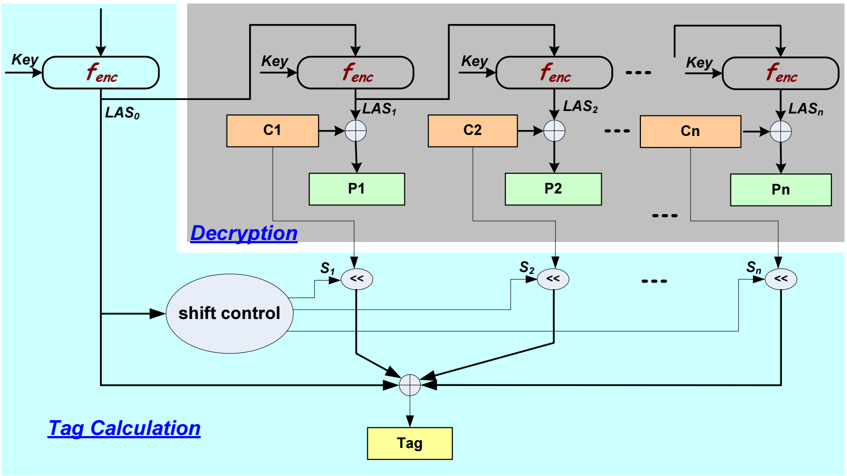


Fig. 4. Decryption and Tag Calculation

Unlike for the memory write operation, where the input plaintext is immediatelyavailable from the cache for encryption, a long latency time is often needed to obtain the ciphertext for a memory read operation. Therefore, the parallel execution is different from that in the memory write operation, as illustrated in Figure 5. As can be seen from Figure 5, the memory latency can be used to perform the time-costly encryption function, f_{enc}; therefore, only the XOR function added to the delay to the critical path for the cache read operation.

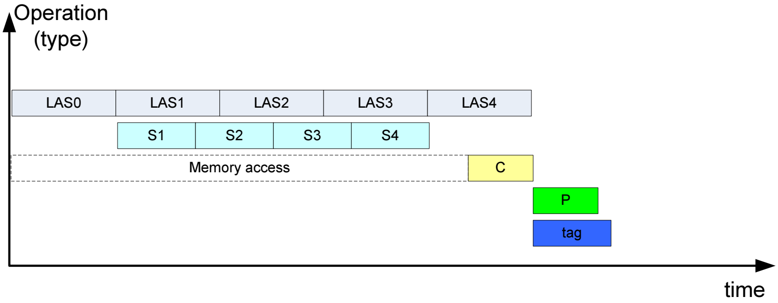
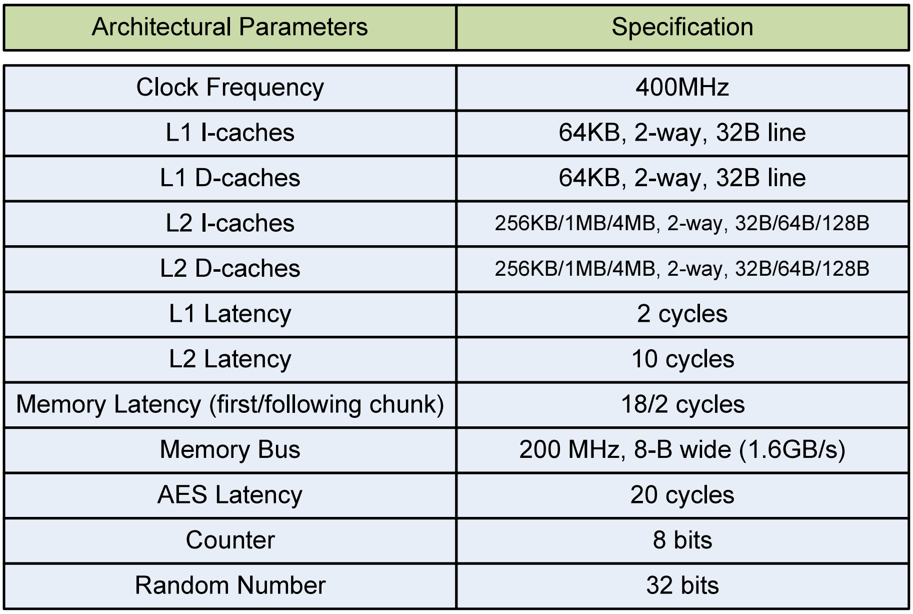


Fig. 5. Parallel Operations in Decryption and Tag Calculation During Memory Read

4 Experimental Results

To evaluate our design, we have built a simulation environment based on the SimpleScalar simulation suite[15]. We use the speculative out-of-order simulator with PISA instruction set architecture. A memory hierarchy that contains two-level instruction and data caches is applied in the architecture.

The SimpleScalar is modified to implement on-chip *FEDTIC* functions. The baseline design architectural parameters used in the simulation are shown in Figure 6.



Architectural Parameters	Specification
Clock Frequency	400MHz
L1 I-caches	64KB, 2-way, 32B line
L1 D-caches	64KB, 2-way, 32B line
L2 I-caches	256KB/1MB/4MB, 2-way, 32B/64B/128B
L2 D-caches	256KB/1MB/4MB, 2-way, 32B/64B/128B
L1 Latency	2 cycles
L2 Latency	10 cycles
Memory Latency (first/following chunk)	18/2 cycles
Memory Bus	200 MHz, 8-B wide (1.6GB/s)
AES Latency	20 cycles
Counter	8 bits
Random Number	32 bits

Fig. 6. Architectural Parameters

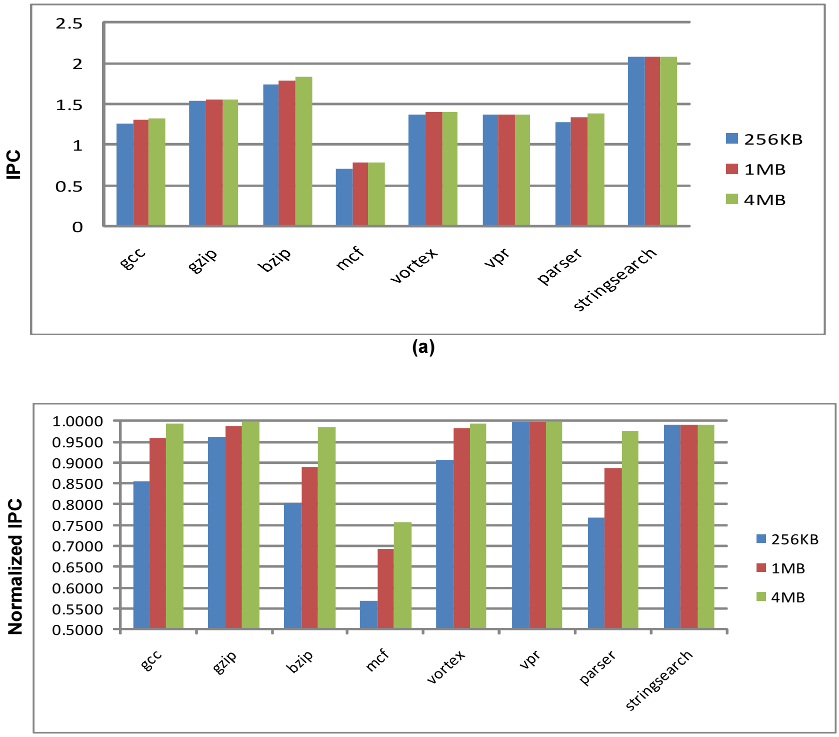
We chose AES as our encryption function. The Xilinx FPGA implementation of AES presented in [16] was integrated in our design. Seven SPEC2000 [17] CPU benchmarks and one MiBench[18] benchmark ($stringsearch$) were used in the simulation.

In our simulation parameters, the processor speed is 400MHz and the *FEDTIC* operates on a FPGA at 200MHz. Therefore, every FPGA computation cycle is equivalent to two processor cycles. The external bus and off-chip memory is assumed to run at 200MHz.

We evaluate the performance overhead from the security design compared to the baseline system without the security protection engine.

The overhead is incurred when a cache miss occurs. On a read miss, the requested cache line is fetched from the external memory, decrypted and tag-checked; on a write miss, the cache line to be written to the external memory is encrypted and tagged. Thus, the overall performance penalty is affected by the cache miss rate, and cache miss penalty. Different cache configurations (cache size, cache line size) affect cache miss rate. We design our experiments with different cache configurations in order to observe the performance overheads. We add a fixed penalty for each memory fetch.

Figure 7(a) shows the baseline performance for each of the benchmarks with different cache sizes (256KB, 1MB, and 4MB) and the same 64Byte cache line. The performance is measured in Instruction Per Cycle (IPC).



(a)

(b)

Fig. 7. Effect of Different Cache Configuration (a) Baseline Performance (b) Performance Overheads

Figure 7(b) illustrates the impact of security protection on the run-time program performance, where the IPC is normalized to the baseline for an easy comparison.

From Figure 7(a), we can see that with increasing cache sizes, the performance is slightly improved, by 2.10% on average. Figure 7(b) shows that the security protection results in some performance overhead that is inversely related to the cache size within the range of 256KB-4MB. With the 256KB-cache, the overhead is up to 43.14% (the worst case from the mcf application) and averaged around 14.36%. When the cache size is increased to 1MB, the average overhead is reduced to 7.66%, and the worst case is improved, to 30.72%, and for some benchmarks, the overhead is less than 5%. When the cache size is further increased to 4MB, the average overhead is only 3.84%, and the worst case is 24.29%.

Table 1 shows the memory and performance overheads for each of the applications when the system has a $1MB$ cache with the cache line of $64Bytes$. The on-chip and off-chip memory costs for each of the benchmarks are given in columns 2 & 4, the percentage values as compared to the baseline design are listed in columns 3 & 5. The IPC values for the baseline design and the design with *FEDTIC* are presented in columns 6 & 7, and the performance overhead as compared to the baseline design is given in the last column (column 8).

Table 1. Memory and Performance Overheads

benchmarks	memory overhead				performance overhead		
	on-chip	%	off-chip	%	baseline (IPC)	FEDTIC (IPC)	%
gcc	4KB	0.39%	173KB	6.24%	1.3130	1.2577	4.21%
gzip	5KB	0.49%	37KB	6.28%	1.5529	1.5348	1.17%
bzip	1KB	0.10%	15KB	6.28%	1.7880	1.5923	10.95%
mcf	0.3KB	0.05%	9KB	6.43%	0.7805	0.5407	30.72%
vortex	2KB	0.20%	69KB	6.23%	1.3954	1.3684	1.93%
vpr	0.6KB	0.06%	25KB	6.25%	1.3645	1.3643	0.01%
parser	4KB	0.39%	30KB	6.25%	1.3451	1.1938	11.25%
stringsearch	0.4KB	0.04%	7KB	6.54%	2.0746	2.0525	1.07%
average		0.21%		6.25%			7.6%

From the simulation results, we can see that our security design incurs little overheads, on average, about 6.25% on the off-chip memory, 7.6% on the system performance, and only 0.21% on the on-chip memory, for the design with a $1MB$ cache and the cache line of $64Bytes$.

5 Conclusion

In this paper, we presented an efficient encryption/authentication scheme to protect the confidentiality and integrity of the data that are processed in a system which has a secure on-chip computing environment and insecure off-chip memory. Our design is easy to implement. We have modeled our design based on the Simplescalar simulation software. Our experiment on a set of benchmarks demonstrates that our security design incurs very little on-chip memory consumption, about 0.21%, and the overheads on the off-chip memory and the system performance are only 6.25% and 7.6%, respectively.

References

1. Gelbart, O., Leontie, E., Narahari, B., Simha, R.: A compiler-hardware approach to software protection for embedded systems. Computers and Electrical Engineering, 315–328 (2009)
2. Ravi, S., Raghunathan, A., Chakradhar, S.: Tamper resistance mechanisms for secure embedded systems. In: 17th International Conference on VLSI Design (2004)
3. Best, R.M.: Prevent software piracy with crypto-microprocessors. In: IEEE Computer Society International Conference (1980)
4. Dallas Semiconductor (2008), http://www.maximic.com/Microcontroller.cfm
5. Blum, M., Evans, W., Gemmell, P., Kannan, S., Naor, M.: Checking the correctness of memories. In: 32nd Annual Symposium on Foundations of Computer Science (1991)
6. Lie, D., Chandramohan, T., Mitchell, M., Lincoln, P., Boneh, D., Mitchell, J., Horowitz, M.: Architectural support for copy and tamper resistant software. In: 9th Internatinal Conference Architectural Support for Programming Languages and Operating Systems, ASPLOS-IX (2000)

7. Lie, D., Thekkath, C.A., Horowitz, M.: Implementing an untrusted operating system on trusted hardware. In: 19th ACM Symposium on Operating System Principles (2003)
8. Suh, G.E., Clarke, D., Gasend, B., van Dijk, M., Devadas, S.: AEGIS:architecure for tamper-evident and tamper-resistant processing. In: International Conference on SuperComputing (2003)
9. Suh, G.E., óDonnell, C.W., Sachdev, I., Devadas, S.: Design and implementation of the AEGIS single-chip secure processor using physical random functions. In: 32nd Interntional Symposium on Computer Architecture, ISCA (2005)
10. Suh, G.E., óDonnell, C.W., Sachdev, I., Devadas, S.: AEGIS: A single-chip secure processor. IEEE Design and Test of Computers, 467–477 (2007)
11. Duc, G., Keryell, R.: Cryptopage: An efficient secure architecture with memory encryption, integrity and information leakage protection. In: 22nd Annual Computer Security Applications Conference, ACSAC 2006 (2006)
12. Elbaz, R., Champagne, D., Lee, R.B., Torres, L.: Tec-tree: a low-cost, parallelizable tree for efficient defense against memory replay attacks. In: Paillier, P., Verbauwhede, I. (eds.) CHES 2007. LNCS, vol. 4727, pp. 289–302. Springer, Heidelberg (2007)
13. Bellare, M., Namprempre, C.: Authenticated encryption: Relations among notions and analysis of the generic composition paradigm. Journal of Cryptology 21(4), 469–491 (2008)
14. Rogaway, P., Bellare, M., Black, J., Krovetz, T.: OCB: a block-cipher mode of operation for efficient authenticated encryption. In: ACM conference on Computer and communications Security (2001)
15. Austin, T.M., Burger, D.B.: The simplescalar tool set, version 3.0. Technical report, University of Wisconsin-madision (1997)
16. Helion Technology Datasheet: high performance AES (Rijndael) cores for Xilinx FPGAs (2008), http://www.heliontech.com
17. Henning, J.L.: SPEC CPU 2000: Measuing CPU performance in the new millennium. IEEE Computers (2000)
18. Guthaus, M.R., Ringenberg, J.S.: Mibench: a free, commercially representiative embedded benchmark suite. In: IEEE 4th Annual Workshop on Workload Characterization (2001)

Enhanced Technique for Secure Wireless Sensor Routing with Respect to Energy Conservation

Maqsood Mahmud[1,3,4], Abdulrahman Abdulkarim Mirza[1,4], Ihsan Ullah[2],
Naveed Khan[2], Abdul Hanan Bin Abdullah[3], and Mohammad Yazid Bin Idris[3]

[1] Department of Information System
[2] Department of Computer Science,
College of Computer and Information Sciences
King Saud University, Saudi Arabia
[3] Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia, Malaysia
[4] Center of Excellence in Information Assurance (CoEIA),
King Saud University, Saudi Arabia
{maqsood.m,amirza,ihsanullah,naveed}@ksu.edu.sa,
{hanan,yazid}@utm.my

Abstract. This paper presents a routing protocol architecture based on recursive group algorithm. This algorithm apply Group Verification Tree approach which makes the sensor network secure and make it safer from malicious intrusions and illegitimate users. The proposed approach will give a new dimension to the fast and secure routing in the sensor networks with less energy to be consumed. Based on the analysis and simulation the proposed strategies yield better results than the existing results.

Keywords: Group Verification Tree, Malicious, Routing Protocol, Wireless Sensor Networks.

1 Introduction

Our introduction is based on the following measurements and concepts.

1.1 Sample Attacks on Routing

The following attacks are brought into account while studying the above algorithm [5][9]. Inject incorrect routing information or alter setup/update messages like Compromised sensors are most problematic. It provides malicious routing data/messages suppress (selectively) routing messages. Specific attacks are Black hole, Wormhole, Replication and Denial of Service.

1.2 Techniques for Secure Routing

The basic three techniques that we are using for secure routing are Prevention, Detection & Recovery and Resilience [3] which uses certain techniques discussed later.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 376–382, 2010.

2 Related Work

The Sensor routing is the most assumed trusted environment. INSENS is only applicable to certain topologies. SIGF requires GPS and Other secure routing protocols [7]. It typically relies on a single technique. For Prevention it uses S-BGP, Ariadne. For Detection & Recovery it uses Watchdog, Pathrater, and Secure Traceroute. While Resilience uses INSENS[8]. The inappropriate resource-constrained sensor nodes require PKI or excessive amounts of memory, computation or communication [13]. Wireless Sensor Networks technology becomes progressively more valuable in public safety, home, medical and office security as well as in military security. Secure-SPIN implements in wireless sensor Networks has three phases using of PASC protocol for confidentiality, eliminating of malicious user through Hash function and in energy conservation through CDMA code [1]. Energy consumption is a key measure in sensor Networks. The multiple node-disjoints paths can be discovered through distributed multi path routing protocol and energy a wearing routing protocols. The load balancing algorithm is used to distribute traffic over multiple paths [2].

3 Routing Protocol Architecture Used

This paper establishes routing tables and network addresses using prevention techniques to thwart active attackers. To detect and recover from attempts to deviate from the protocol or to launch additional attacks and apply resilient routing techniques to forward packets. It uses the securely established routing tables and network addresses [1].

3.1 Our Assumptions

Our assumptions are Network authority (NA) uses a public/private key pair $\{K_{NA}, K^{-1}_{NA}\}$, each sensor node preloaded with, Network authority's public key K_{NA} ,Unique $ID_{x,}$ Certificate: Sig(K^{-1}_{NA}, ID_x), Signature scheme optimizes for verification, Intended for networks of primarily stationary sensors.

3.2 Address and Route Setup

Goal assigns a unique network address to each node to populate each node's routing table to accomplished with a recursive grouping algorithm, initially, each sensor constitutes its own group, groups repeatedly merge until all nodes belong to same group. Each time a node's group merges, the node adds one bit to its network address and one entry to its routing table.

3.3 Recursive Grouping Algorithm

In this scenario each Group act in an asynchronous, distributed fashion and is explained below categorically [12]. Each group collects information about its neighbors, proposes to merge with smallest neighboring group. It is based on number of nodes in the group, ties broken based on group ID. This metric keeps addresses and routing tables small. The mutual proposal triggers merge entire process is deterministic for a given topology. It limits the damage. An attacker can inflict.

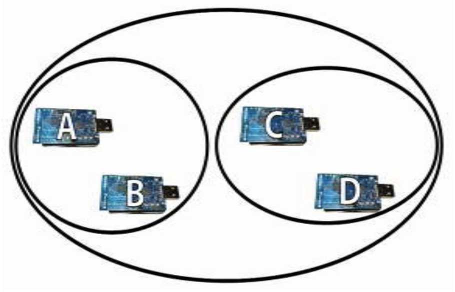


Fig. 1. Initial Convergence

Table 1. Detecting Grouping Deviations

Node Id	Address	Routing Table
A	0.0	RT_A
B	0.1	RT_B
C	1.0	RT_C
D	1.1	RT_D

3.4 Calculating Network Addresses

Assume G and G' decide to merge each node in G independently extends its network address by one bit. It is based on Nodes in G' which make similar changes while merging.

$$R_i = \begin{cases} 0 & ID_G < ID_{G'} \\ 1 & ID_G > ID_{G'} \end{cases} \quad (1)$$

3.5 Populating Routing Tables

Let's assume G and G' decide to merge and each node in G records the neighbor from whom it heard about G' in its current routing table slot

3.6 Forwarding

In this method the basic forwarding is similar to area-style forwarding. If given a destination network address route towards node with longest matching prefix will be adopted. (i) Path length in logical hops bound by log (n), (ii) A logical hop may require several physical hops.

4 Threats

The compromised nodes may lie about group size or ID to subvert route setup and compromised nodes may claim multiple IDs or try to simultaneously group with several other nodes [2][9][10].

4.1 Detecting Grouping Deviations

To Maintain a Grouping Verification Tree (GVT) for each group during recursive grouping it prevents attacker from lying about group ID or size, based on a hash tree construction [6]. Before two groups merge, they verify each other's GVT. Integrity of the GVTs insures integrity of the recursive grouping algorithm. Final GVT covers all nodes in the network. It can be used to authenticate any node's network address to prevent illegal node to the sensor networks.

4.2 Hash Trees

We used Hash function which has O(1) time complexity. To employ a one-way hash function H: {0,1}*→{0,1}ρ has to be observed. To create a one-way data structure-the Merkle Tree is one such data structure that has to be used.

- Each internal node calculated as:
  Parent = H(ChildL || ChildR)
- Authenticates a leaf node given the root value and nodes along the path to the root

4.3 Group ID Computation

To assume G and G' decide to merge. Each node in G independently calculates the new group ID as:

$$ID_\gamma = \begin{cases} h(ID_G, |G|, ID_{G'}, |G'|) & ID_G < ID_{G'} \\ h(ID_{G'}, |G'|, ID_G, |G|) & ID_G > ID_{G'} \end{cases} \tag{2}$$

4.4 GVT Formation

There will be one GVT per group. The GVT leaves are IDs of nodes in the group. Internal nodes represent intermediate group IDs. Each node maintains information about its branch of the GVT specifically, the group ID and size of each merge partner.

4.5 GVT Verification

Before merging, group G verifies the GVT for G' (and vice versa). G' announces its group ID (and size). Group G sends a challenge value to G'. The challenge uniquely selects a node in G'. Chosen node sends its certificate and GVT information to G. Nodes in G verify the GVT values [11]. By this mechanism all are verified to be non malicious nodes. During the convergence of the GVT mechanism the verification process automatically eliminates malicious nodes in the sensor network and the authentication is denied for malicious nodes.

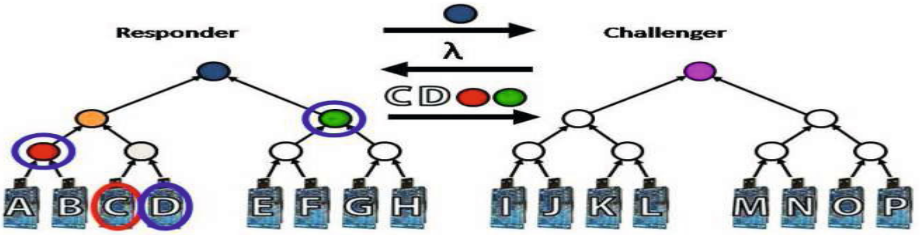


Fig. 2. Response and Challenge Scenario of Sensor Nodes

4.6 Eliminating Malicious Nodes

To legitimate nodes we used the Honeybee mechanism to eliminate malicious nodes. To revoke malicious node M, legitimate node L broadcasts: IDL, IDM, and a signature legitimate nodes revoke M *and* L. It prevents a compromised node from revoking more than one legitimate node.

5 Simulations

We made comparison against Beacon Vector Routing (BVR) protocol [NSDI 2005] with energy saving in mind [8]. It is optimized for efficiency. Experimental Setup for 500 nodes, random deployment, DOI radio model are made to achieve our results. Summary of our Results are:

- Paths longer than shortest path
- Distributes overhead evenly throughout network
    - Better than BVR, even in topologies with voids
- Our routing success rate is 100% as it is a trade between speed and security. It is because of low weight cryptography (Stream Ciphers) and use of Hash Function which has time complexity of O(1).

5.1 Metric: Path Stretch

Stretch = Protocol Path Length / Optimal Path Length.
Optimistic for BVR: does not include failed BVR routes.

5.2 Implementation

We developed in NesC on TinyOS using Telos sensor nodes with source code to be available soon. The challenges overcome with (i) Reliable Broadcast (ii) A synchronicity (iii) Asymmetric Links. Ongoing work to expand the current test bed.

5.3 Validation

Our proposed research work is different and unique from the previous work done on sensor routing. As the verification during the convergence in the GVT mechanism authentication is denied for malicious nodes. Second because of low weight cryptography (stream ciphers) less computational time is required due to which less energy is required for sensor routing networks. More over our solution is more valid due to usage of Hash Function which has computational complexity of O(1), which is fastest for measuring the speed of convergence.

6 Critiques and Future Work

Following critiques and suggestions came after thorough analysis of relevant papers.

1. This paper used Tree for verification of nodes and emerged with algorithm i.e. **GVT** (Group Verification Tree), so that illegitimate sensor devices could not enter into the network as an intruder. This algorithm can be made more enhanced and efficient by using Hash function rather than trees. This will exponentially increase its routing information convergence process with less minimum runtime if memory is provided to maintain hash table.
2. As for as security is concerned, sensor device may be incorporated and embedded with a chip with secure encrypted ID .This encryption can be done using any Asymmetric or Symmetric cryptography(Public Key Cryptography). I proposed stream cipher to be more suitable for this system because of light weight cryptography and good avalanche effect of sensors and wireless networks.
3. The secure routing protocol should be introduced to combat the hidden terminal problem.
4. The protocol should use (less control messages).e.g. RREQ. (Route Request)
5. The computational power of the algorithm should be less to save the sensor *Node Energy* [4]
6. If the above suggestions are brought into consideration, then IEEE Standard 802.11.15.4 can be further improved and enhanced.

7 Conclusions

To secure sensor routing is an important and difficult problem. Most previous techniques assume a trusted environment or use a single security technique. The authors designed a protocol incorporating all three security techniques that still compares favorably to insecure protocols.

Acknowledgment

We are thankful to Chairman of Department of Information Systems, Dean of College of Computer and Information Sciences, Center of Excellence in Information Assurance, King Saud University for their cooperation and support.

References

1. Xiao, D., Wei, M., Zhou, Y.: Secure-SPIN: Secure Sensor Protocol for Information via Negotiation for Wireless Sensor Networks. In: 1ST IEEE Conference on Industrial Electronics and Applications, pp. 1–4 (2006), doi:10.1109/ICIEA.2006.257149
2. IEEE 64th Vehicular Technology Conference VTC 2006, pp. 1–5 (Fall 2006), doi:10.1109/VTCF.2006.505
3. Akyildiz, S., et al.: Wireless Sensor Networks: A Survey (2002)
4. Ganesan, D., et al.: Highly Resilient, Energy-Efficient Multipath Routing in Wireless Sensor Networks. Mobile Comp. and Commun. Review 5(4), 10–24 (2002)
5. Hu, Y.-C., Perrig, A., Johnson, D.B.: Packet Leashes: A Defense Against Wormhole Attacks in Wireless Networks. In: Proc. IEEE INFOCOM (2003)

6. Parno, B., Luk, M., Gaustad, E., Perrig, A.: Secure Sensor Network Routing: A Clean Slate Approach (2006)
7. Chan, H., Perrig, A., Song, D.: Random Key Predistribution Schemes for Sensor Networks. In: Proceedings of the 2003 IEEE Symposium on Security and Privacy, May 11-14, p. 197 (2003)
8. Tejaswi, K., Mehta, P., Bansal, R., Parekh, C., Merchant, S.N., Desai, U.B.: Routing Protocols for Landslide Prediction using Wireless Sensor Networks (2006)
9. Roy, S., Setia, S., Jajodia, S.: Attack-resilient hierarchical data aggregation in sensor networks. In: Proceedings of the fourth ACM workshop on Security of ad hoc and sensor networks, Alexandria, Virginia, USA, October 30 (2006)
10. Chan, H., Perrig, A., Przydatek, B., Song, D.: SIA: Secure information aggregation in sensor networks. Journal of Computer Security 15(1), 69–102 (2007)
11. Srinivasan, A., Wu, J.: A novel k-parent flooding tree for secure and reliable broadcasting in sensor networks. In: Proceedings of IEEE International Conference on Communications—Computer and Communications Network Security, ICC CCN (2007)
12. Srinivasan, A., Wu, J.: Secure and reliable broadcasting in wireless sensor networks using multi-parent trees. In: Security Comm. Networks, Wiley InterScience, Hoboken (2008)
13. Roman, R., Lope, J.: Integrating wireless sensor networks and the internet: a security analysis. Journal of Internet Research 19(2), 246–259 (2009)

End-to-End Security Methods for UDT Data Transmissions

Danilo Valeros Bernardo and Doan B. Hoang

iNext, Computing and Communications
Faculty of Engineering and Information Technology
The University of Technology Sydney, Sydney
Australia
bernardan@gmail.com, dhoang@it.uts.edu.au

Abstract. UDT (UDP-based data transfer protocol) is one of the most promising network protocols developed for high data speed data transfer. It does not, however, have any inherent security mechanisms, and thus relies on other transport protocols to provide them. Towards its implementation in high speed networks, security and privacy are critical factors and important challenges that need to be addressed. There were substantial research efforts we carried out so far to address these challenges. We introduced security mechanisms through the application layer using UDT's API and presented DTLS, GSS-API, and CGA, in transport and IP layers. In this paper, we make the following contributions: we out line security requirements for UDT implementation and propose practical encryption methods for securing UDT within the network layer.

Keywords: UDT, TCP, GSS-API, DTLS, Next Generation Protocol.

1 Introduction

Developments in 2007 introduced UDT, the next generation of high performance data transfer protocol for cloud computing [3], [4]. One compelling example of the implementations of UDT is the Sloan Digital Sky Survey (SDSS) project [3-6], which is mapping in detail one quarter of the entire sky, determining the positions and brightness of more than 300 million celestial objects. It measures the distances to more than a million galaxies and quasars. The data from the SDSS project so far has increased to 2 terabytes and continues to grow. Currently, the 2 terabytes data is being delivered to the Asia-Pacific region, including Australia, Japan, South Korea, and China. Astronomers also want to execute online analysis on multiple datasets stored in geographically distributed locations [3-6].

This implementation offers a promising direction for future high speed data transfer in various industries.

The absence of a well-thought security mechanism for UDT when it was developed, however, drives this paper to introduce ways to secure UDT in various environment and implementation scenarios.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 383–393, 2010.

In the following sections, we briefly introduce UDT. We then present the objectives of this paper; introduce and discuss the methods in securing UDT and present the results of experiments, and present the conclusion of this paper.

1.1 Background

UDT introduces a new three-layer protocol architecture that is composed of a connection flow multiplexer, enhanced congestion control and resource management. The new design allows protocol to be shared by parallel connections and to be used by future connections. It improves congestion control and reduces connection set up time. UDT provides better usability by supporting a variety of network environments and application scenarios [6]. It addresses TCP's limitations by reducing the overhead required to send and receive streams of data.

UDT is a connection-oriented duplex protocol, which supports data streaming and partial reliable messaging [3-6], [12]. It also uses rate-based congestion control (rate control) and window-based flow control to regulate outgoing traffic. This was designed such that rate control updates the packet sending period every constant interval, whereas flow control updates the flow window size each time an acknowledgement packet is received. It was expanded to satisfy more requirements for both network research and applications development. This expansion is called Composable UDT and was designed to complement the kernel space network stacks.

The pressure, however, to reduce the cost and complexity of running streaming applications over the Internet and through wireless and mobile devices continues to mount.

Moreover, users demand better security and privacy for their communication links.

Our contention for the need of security mechanisms of the new UDT is derived from 5 important observations [3-6].

- Absence of inherent security mechanism, such checksum for UDT.
- The header information is not sufficient for this protocol.
- Dependencies on user preferences and implementation on the layer on where it is implemented.
- Dependencies on existing security mechanisms of other layers on the stack.
- Dependencies on TCP/UDP which are dependent on nodes and their addresses for high speed data transfer protocol leading to a number of attacks such as neighborhood, Sybil and DoS (Denial of Service) attacks.

Earlier works in the development of security framework for UDT support the need of minimizing its sending rates [3], [6] in retransmissions and introducing its own checksum in its design. The introduction of other security mechanisms, however, to secure UDT is presented to address its vulnerabilities against adversaries exploiting the application, transport, and IP layers.

We [3-6] presented an overview on securing UDT implementations in various layers. However securing UDT in application and other layers need to be explored in future UDT deployments in various applications.

There are application and transport layer based authentication and end-to-end [7] security options for UDT.

1.2 Motivations

In this paper, there are 2 important objectives we intend to address:

Firstly, reviews on existing methods of security feasible for UDT transmission, such as:

- Security at the application layer via UDT extensions may require client and servers, and significant changes on applications to accommodate security features.
- Encryption be performed at the layer 3 (Network Layer), abstracted from the UDT application, e.g., via gateway-to-gateway, virtual private networks (VPNs), when security solution on the application layer becomes too complex to develop.

Secondly, where sessions are point-to-point we propose the use of host-to-host encryption, or gateway-to-gateway encryption at the border of each host's network. Utilizing encryption security, where the absence of a viable security mechanism of a particular new protocol such as UDT is a feasible option.

We looked at existing encryption support for UDP and TCP, and determined no encryption methods were proposed and available for UDT. However, our assertion is that encryption can support UDT through the network layer when it is running on top of UDP.

Existing security mechanisms for UDP developed at some layers are certainly not advanced and flexible to operate with UDT, e.g., UDP: UDT+DTLS and UDT+GSS-API, UDT+MD5, SHA-1 or 256 options as proposed earlier in our works [6]. The progression of these mechanisms require significant changes on the UDT's design structure, such as accommodating large APIs and functions for the development security mechanisms; and introducing an option for hash functions to integrate in its header to secure network gateway connections. We present in this paper that existing encryption methods particularly on the network layer used for other protocols can simply be implemented on UDT and UDP while the development of proprietary mechanisms is underway.

2 Encryption Methods

We describe 2 encryption methods not specific to UDT in which it can operate. These methods can be used to secure communications between networks, for any application using different protocols.

First, host-to-host method. In this method the encryption is set up between the two hosts wishing to exchange secure data. This has the advantage that encryption is applied end-to-end along with the full path of the connection, but it also implies the user may have to do something to initiate the encryption.

Second, gateway-to-gateway method, which encryption is applied between boundary routers at the edge of a network. This means that sessions are encrypted end-to-end, and thus are not vulnerable to snooping within the endpoint networks.

In the experiments presented in section 3, we report on tests performed on one method (gateway-to-gateway) rather than Host-to-Host method when UDT is implemented on top of UDP.

Host-to-Host is simplistic and can be attained by implementing Windows built in IPsec. This method is attainable through implementing existing OS and applications that provides end-to-end security. Subsequently, our focus on gateway-to-gateway method is relevant due to the behavior of UDT in data transmission where it performs significantly long distance high bandwidth networks (via the Internet) which most of them are behind gateways.

2.1 Internet Protocol Security (IPsec)

IPsec offers the promise of protecting against many of denial-of-service attacks. It also offers other potential benefits. Conventional software-based IPsec implementations isolate applications from developing proprietary cryptographic keys, improving security by making inadvertent or malicious key exposure more difficult. In addition, specialized hardware may allow encryption keys protected from disclosure within trusted cryptographic units. Also, custom hardware units may well allow for higher performance [6].

Implementing UDT running with IPsec provides adequate protection for data transmission (fig. 1). A datagram-oriented client application using UDT will use the connection-oriented part of its API (because it is using a given datagram socket to talk to a specific server) while the server it is talking to use the connection oriented API because it is using a single socket to receive requests from and send replies to a large number of clients.

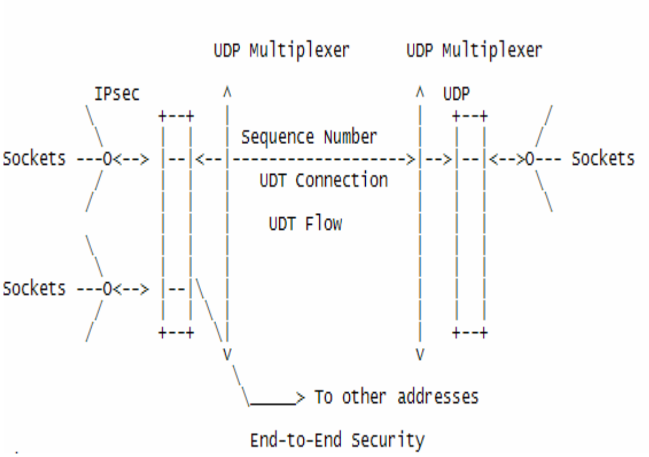


Fig. 1. UDT flow using end-to-end security. [8-11]. IPsec can be used without modifying UDT and its applications running it.

2.2 IPsec Tools (Gateway-to-Gateway Method)

There are a number of commercial IPsec VPN solutions on the market, of which Netscreen is a typical example. The key management scheme is IKE, Sha, Key Group Diffie-Hellman Group 2, Encryption algorithm is Triple DES.

Configuration on the Netscreen firewall for keys and a client software needs to be configured on the host machine. However, since we are looking at gateway-to gateway encryption, we use the site-to-site VPN configuration.

The product also provides secure remote access to a corporate network, client- to-site (host –to-gateway), and is able to scale up to multiple remote sites.

We also consider a downloadable IPsec software for implementation running on Linux platform.

This software is Free Secure Wide Area Network (S/WAN) or FreeS/WAN that can be used with PGP and X .509 certificates. It uses IKE for key exchange. It has an important feature called "opportunistic encryption" such that any two FreeS/WAN gateways will encrypt data when traffic is observed flowing between them.

This software, however, has ceased for further developments due to export and legal restrictions in the United States. However, its last version remains to be free and suitable for non-commercial use, such as research. We selected this tool because of its flexibility and its interesting feature of opportunistic encryption. It also has very low overhead compared to using other software.

We tested UDT on both tools. In section 3, we present the results of our experiments using Netscreen VPN and FreeS/WAN. A simple application running UDT was downloaded [3] and installed on Windows OS hosts behind the secure gateways.

3 Encryption Experiments

We describe and conduct experiments on gateway-to-gateway encryption method under two environments.

In our tests, we evaluated UDT performance using FreeS/WAN in gateway-to-gateway mode. We also tested UDT using a commercial product called Netscreen in site-to-site mode. For FreeS/WAN, the test network composed of two gateway hosts. For Netscreen, we used two firewalls at the gateways and configured them for site-to-site VPN. The client PCs used for the test were a 32GB 2.3GHz laptop running Windows XP Service Pack 2, with an application running UDT and they were behind the gateways.

3.1 FreeS/WAN (Gateway-to-Gateway)

We summarize the tests performed on FreeS/WAN. We carried out tests on FreeS/WAN in tunnel mode (gateway-to-gateway) rather than in transport mode (i.e. host-to-host).

The default algorithms used by FreeS/WAN are 3DES and MD5, with RSA authentication keys. These are defined in the */etc/ipsec.conf* file. Private and other key information is stored in */etc/ipsec.secrets*.

3.2 Netscreen IPsec VPN (Site-to-Site)

We configured the firewalls site-to-site VPN and allowed traffic from a selected gateway private network. We created route-based VPNs and bind their configuration to a virtual interface called tunnel interface, with fixed IP addresses. We assigned address for the tunnel interface, with both firewalls that make up the tunnel using route-based, in the same subnetwork. We created the IKE gateway for Phase-1 which the same for route and tunnel interface and created autokey IKE and bind it to the interface. We then configured the routing to the remote network to the outbound interface.

3.3 Measurement Schemes

To test the performance, we repeated ping and ftp tests using existing performance measurement tools. For ping and throughput measurements we used Qcheck/Statscout, and MS' pathping. We also used WS_FTP server (running on the laptop device), see results on tables 3 and 4.

Qcheck/Statcout allows network performance measurements to be taken between any two endpoints, with control from a remote console program. Qcheck results were as follows, averaging over 10 measurements see tables 1 and 2):

Table 1. Results of Qcheck, with and without encryption

Netscreen (10repeats)

Measurement	Average Plain	Average with Encryption
TCP throughput	200.9 ms	180.4 ms
UDP throughput	140.2 ms	130.3 ms
UDT throughput	9000 ms	6000 ms
TCP response time	1/1/1 ms min/avg/max	.8/1/1.2 ms min/avg/max
UDP response time	1/1/1 ms min/avg/max	1/1/1 ms min/avg/max
UDT response time	1/1/1 ms min/avg/max	.9/.95/1 ms min/avg/max

Table 2. Results of Qcheck, with and without encryption

FreeS/WAN (10repeats)

Measurement	Average Plain	Average with Encryption
TCP throughput	187 ms	162 ms
UDP throughput	120.1 ms	113 ms
UDT throughput	7000 ms	4000 ms
TCP response time	1/1/1 ms min/avg/max	1.2/2.1/3 ms min/avg/max
UDP response time	1/1/1 ms min/avg/max	1.2/2.1/3 ms min/avg/max
UDT response time	1/1/1 ms min/avg/max	1.2/2.1/3 ms min/avg/max

These results do not suggest there would be problems encrypting the sessions of UDP and UDT.

Table 3. FTP test results with (encrypted data) and without (or plain data) encryption

FTP (Netscreen)

FTP Passive Mode	Time (sec)	Plain Mbytes	Time (sec)	Encrypted Mbyte/s	Repeat
Laptop-gateway-gateway-laptop	2.3	1.2	1	0.6	1
	2	1.3	1	0.8	2
	1.8	1	1	0.6	3
	2.2	1.2	1	0.7	4
	2.4	1.2	2	0.8	5
	1.9	1.05	1	0.5	6
	2	1.3	1	0.5	7
	2.6	1.4	2	1	8
	2	1.4	1	0.8	9
	1.8	1.2	1	0.9	10

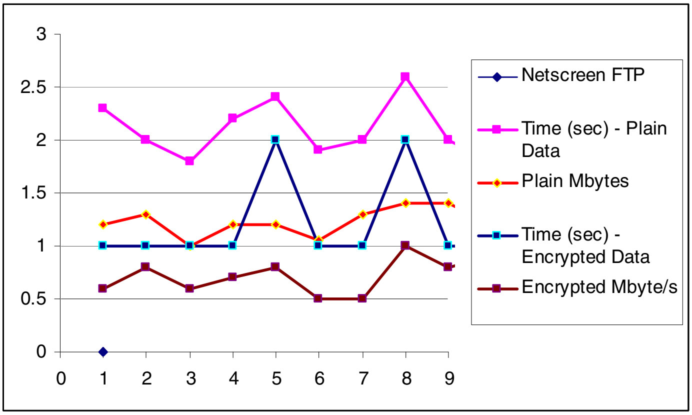


Fig. 2. Netscreen Graph results

The encryption on our test network was performed on the higher specification gateways and firewalls.

The resulting performance is acceptable with encryption. The significant result is that the encrypted FTP throughput is seen to be acceptable with 90% of the non-encrypted throughput (see tables 3 and 4, figs 2 and 3). This reflects the use of high speed bandwidth.

Finally, we performed *ping* tests (with WS Ping Pro and pathping) between the two gateways with differing size *ping* packets, with the response time in milliseconds being measured. We also ran the ping tests within the interfaces of the firewalls.

Table 4. FTP test results with (encrypted data) and without (or plain data) encryption

FTP (FreeS/WAN)

FTP Passive Mode	Time (sec)	Plain Mbyte/s	Time (sec)	Encrypted Mbyte/s	Repeat
Laptop-gateway-gateway-laptop	2.1	1.2	2	0.6	1
	1.6	1.3	2	0.8	2
	1.2	1	2	0.6	3
	1.8	1.2	1	0.7	4
	2	1.2	1	0.8	5
	2.2	1.05	1	0.5	6
	1.8	1.3	1	0.5	7
	2.1	1.4	2	1	8
	1	1.4	1	0.8	9
	1.8	1.2	1	0.9	10

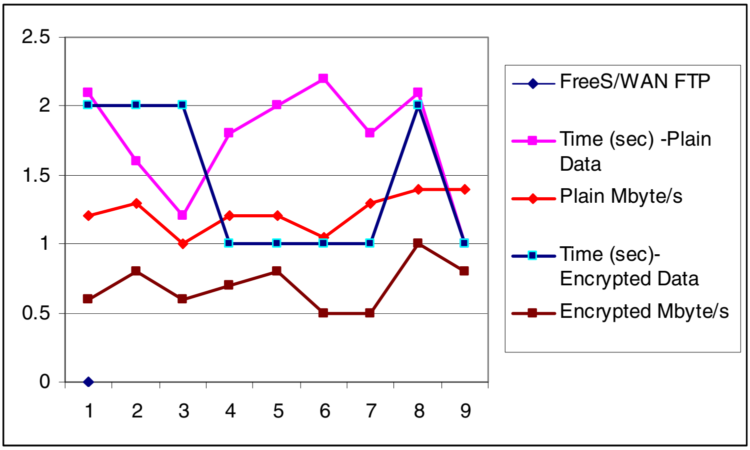


Fig. 3. FreeS/WAN Graph results

Table 5. Ping test results, with and without encryption

Netscreen (Ping test)

Gateway	Mode	1000 bytes	2000 bytes	10000 bytes
Laptop-gateway-gateway laptop	Clear	1000	1988	9800
	IPsec	1000	1990	10000

Table 6. Ping test results, with and without encryption

FreeS/WAN (Ping test)

Gateway	Mode	1000 bytes	2000 bytes	10000 bytes
Laptop-gateway-gateway laptop	Clear	1000	1900	9600
	IPsec	1000	1800	10000

The raw response times are poorer (see tables 5 and 6) for our non-encrypted (clear) test at the default *ping* packet size, but we presume the intervening gateways are responsible for that effect. Otherwise, the results are significantly improved, especially for the much larger packet sizes where fragmentation is occurring.

Overall our results illustrate the improved performance that using higher specification encrypting devices with higher bandwidth links can bring. The various throughput and ping tests we performed suggest that at the bandwidth levels required fast data transfer, Netscreen would appear to offer a solution that doesn't impact significantly on latency or session quality.

4 Impact on Performance

Our experience with trials of both Netscreen and FreeS/WAN implies that encryption at the network layer does not impose significant performance problems. Perceived latency was very similar in our tests, and the empirical results imply the overhead at the network layer is in the order of a handful of milliseconds.

It should be noted that in addition to the encryption overhead on the CPU, the encrypted packets will also be larger due to the additional AH/ESP data being sent, and the packet re-assembly at the far end will take longer, i.e. there are delays in passing encrypted data beyond just the raw computational burden [7].

In terms of algorithms, one might assume that weaker algorithms are less computationally expensive. However, existing encryption algorithms such as AES, can offer better encryption in comparison to 3DES for less processor effort (an important consideration when encryption is required on miniature smart type devices). Coupled with advances in processor and bandwidth speed, the latency penalty for encryption will continue to fall as a percentage of the time and bandwidth required for high data transfer. The same, of course, may not be true for UDT implementations over a low bandwidth network such as cellular wireless (where the packet size overhead is far more significant). AES also has the advantage of being an open standard [7].

5 Conclusion

There are a number of conclusions that we can draw from the above comparisons and experiments. Since UDT is very new, there is very little adoption of this new protocol, consequently no security mechanisms available for the application layer. We initially consider host-to-host encryption as a feasible solution, depending on the operating system and method desired. However, it is likely to require some expertise in the end user of the end host's system, and will cause some problems for firewalls, because stateful inspection for UDP in which UDT runs on cannot be performed on an encrypted session [7].

We assert that gateway-to-gateway encryption would appear to offer a flexible and relatively efficient means to encrypt UDT and UDP data over the public (Internet) part of a session connection, but is vulnerable to snooping on the internal site network behind the gateway, unless appropriate security solutions are also put in place.

We observed that latency effects of encryption do not appear to be significant based on limited tests performed with FreeS/WAN and Netscreen on entry-level high speed gateway devices. Increasing commodity CPU power is making encryption ever more viable for reasonable UDT data transfer. Opportunistic encryption is desirable. FreeS/WAN includes support for this, but we have been unable to test it in details – the scaling issues may be significant and should be tested further if FreeS/WAN is to be considered for UDT wider deployment. Gateway encryption via a product like FreeS/WAN for smaller and less bandwidth and Netscreen for higher bandwidth and bigger environments may be effective methods at present. Such products and their configurations however, need wider-scale testing prior to their potential use. End-to-end encryption through gateway-to-gateway encryption, therefore offers security for UDT as it is to other protocols.

References

[1] Bellovin, S.: Defending Against Sequence Number Attacks. RFC 1948 (1996)

[2] Bellovin, S.: Guidelines for Mandating the Use of IPsec, Work in Progress, IETF (October 2003)

[3] Bernardo, D.V., Hoang, D.B.: A Conceptual Approach against Next Generation Security Threats: Securing a High Speed Network Protocol – UDT. In: Proc. IEEE the 2nd ICFN 2010, Shanya, China (2010)

[4] Bernardo, D.V., Hoang, D.B.: Security Requirements for UDT, IETF Internet-Draft – working paper (September 2009)

[5] Bernardo, D.V., Hoang, D.B.: "Network Security Considerations for a New Generation Protocol UDT. In: Proc. IEEE the 2nd ICCIST Conference 2009, Beijing, China (2009)

[6] Bernardo, D.V., Hoang, D.B.: A Security Framework and its Implementation in Fast Data Transfer Next Generation Protocol UDT. Journal of Information Assurance and Security 4(354-360) (2009) ISN 1554-1010

[7] Chown, T., Juby, B.: Overview of Methods for Encryption of H.323 Data Streams. Technical Paper, University of Southampton (March 2001)

[8] Blumenthal, M., Clark, D.: Rethinking the Design of the Internet: End-to-End Argument vs. the Brave New World. In: Proc. ACM Trans Internet Technology, vol. 1 (August 2001)

[9] Clark, D., Sollins, L., Wroclwski, J., Katabi, D., Kulik, J., Yang, X.: New Arch: Future Generation Internet Architecture, Technical Report, DoD – ITO (2003)

[10] Falby, N., Fulp, J., Clark, P., Cote, R., Irvine, C., Dinolt, G., Levin, T., Rose, M., Shifflett, D.: Information assurance capacity building: A case study. In: Proc. 2004 IEEE Workshop on Information Assurance, pp. 31–36. U.S. Military Academy (June 2004)

[11] Gorodetsky, V., Skormin, V., Popyack, L. (eds.): Information Assurance in Computer Networks: Methods, Models, and Architecture for Network Security. St. Petersburg, Springer, Heidelberg (2001)

[12] Gu, Y., Grossman, R.: UDT: UDP-based Data Transfer for High-Speed Wide Area Networks. Computer Networks 51(7) (2007)

[13] Hamill, J., Deckro, R., Kloeber, J.: Evaluating information assurance strategies. Decision Support Systems 39(3), 463–484 (2005)

[14] H.I. for Information Technology, H. U. of Technology, et al.: Infrastructure for HIP (2008)

[15] Harrison, D.: RPI NS2 Graphing and Statistics Package, `http://networks.ecse.rpi.edu/~harrisod/graph.html`

[16] Jokela, P., Moskowitz, R., Nikander, P.: Using the Encapsulating Security Payload (ESP) Transport Format with the Host Identity Protocol (HIP). RFC 5202, IETF (April 2008)

[17] Kent, S., Atkinson, R.: Security Architecture for the Internet Protocol. RFC 2401 (1998)

[18] Leon-Garcia, A., Widjaja, I.: Communication Networks. McGraw Hill, New York (2000)

[19] Mathis, M., Mahdavi, J., Floyd, S., Romanow, A.: TCP selective acknowledgment options. IETF RFC 2018 (April 1996)

[20] Menezes, A.J., van Oorschot, P.C., Vanstone, S.A.: Handbook of Applied Cryptography. CRC Press, Boca Raton (1997)

[21] NIST SP 800-37, Guide for the Security Certification and Accreditation of Federal Information Systems (May 2004)

[22] NS2, `http://isi.edu/nsna/ns`

[23] PSU Evaluation Methods for Internet Security Technology, EMIST (2004), `http://emist.ist.psu.edu` (visited December 2009)

[24] Rabin, M.: Digitized signatures and public-key functions as intractable as Factorization. MIT/LCS Technical Report, TR-212 (1979)

[25] Rescorla, E., Modadugu, N.: Datagram Transport Layer Security. RFC 4347, IETF (April 2006)

[26] Rivest, R.L., Shamir, A., Adleman, L.M.: A method for obtaining digital signature and public-keycryptosystems. Communication of ACM 21, 120–126 (1978)

[27] Schwartz, M.: Broadband Integrated Networks. Prentice Hall, Englewood Cliffs (1996)

[28] Stewart, R. (ed.): Stream Control Transmission Protocol, RFC 4960 (2007)

[29] Stoica, I., Adkins, D., Zhuang, S., Shenker, S., Surana, S.: Internet Indirection Infrastructure. In: Proc. ACM SIGCOMM 2002 (2002)

[30] Szalay, A., Gray, J., Thakar, A., Kuntz, P., Malik, T., Raddick, J., Stoughton, C., Vandenberg, J.: The SDSS SkyServer - Public access to the Sloan digital sky server data. In: ACM SIGMOD 2002 (2002)

[31] Wang, G., Xia, Y.: An NS2 TCP Evaluation Tool, `http://labs.nec.com.cn/tcpeval.html`

[32] Globus XIO: http://unix.globus.org/toolkit/docs/3.2/xio/index.html (retrieved on November 1, 2009)

[33] Zhang, M., Karp, B., Floyd, S., Peterson, L.: RR-TCP: A reordering-robust TCP with DSACK. In: Proc. the Eleventh IEEE International Conference on Networking Protocols (ICNP 2003), Atlanta, GA (November 2003)

Evading Virus Detection Using Code Obfuscation

Khurram Murad, Syed Noor-ul-Hassan Shirazi, Yousaf Bin Zikria, and Nassar Ikram

National University of Science and Technology (NUST), Islamabad
khurramjarral@gmail.com, noorshirazi@gmail.com,
yusi_2@hotmail.com, dr_nassar_ikram@yahoo.com

Abstract. The conflict between malware authors and analysts is heating up as both are coming up with new armaments in their armory. Malware authors are employing novel sophisticated techniques like metamorphosis to thwart detection mechanisms while security professionals are budding new ways to confront them. In this paper we formally treat diverse mechanisms of making malware undetectable in general and code mutation techniques in particular. We also supported our argument where possible, through different tools and have revealed their outcome. In the end we give our methodology to make any virus undetectable using amalgamation of hex editing and metamorphic techniques.

Keywords: Computer virus, Polymorphism, Metamorphism, Obfuscation, Hex editing, Virus signature.

1 Introduction

A computer virus is a malicious piece of software that modifies other files to inject its code [1]. A virus can change its code on each infection [2]. Virus detection is an uncertain process [2]. Viral mutation techniques are continuously evolving and progressing to evade antivirus algorithms and tools. This resulted in more and more complex virus families of which metamorphic viruses are the most sophisticated one.

Antivirus systems use various detection techniques including signature detection and code emulation to detect malware. Signature based tools look for particular signature while code emulators execute virus in a virtual environment for detection.

To evade signature detection, virus writers continuously change virus using metamorphic techniques while keeping the same functionality. Metamorphic viruses use different code obfuscation techniques to change the structure of the code. These techniques include code reordering through jumps, subroutine permutation, dead code insertion, equivalent instruction substitution, and rearrangement of instruction order (transposition).

To evade code emulation techniques, various anti-emulation techniques have been developed by the malware writers. These include, Entry Point Obscuring (EPO) techniques, decrypting and running code chunk by chunk, using odd instructions that would deceive an emulator, random concealing of decryption, long looping through dead code, multiple encryption layers. Aforementioned techniques have some drawbacks i.e., considerable increase in size of the morphed copy and loss of functionality.

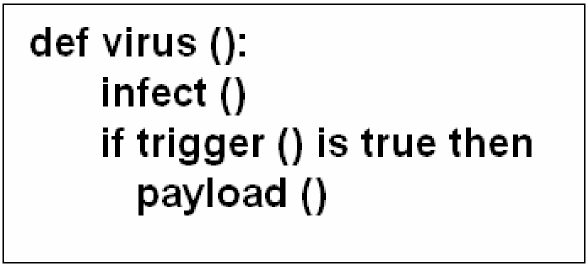
T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 394–401, 2010.

In this work, we focus on making virus undetectable using combination of hex editing and metamorphic techniques to address previous shortcomings.

This paper is organized as follows. Section 2 gives brief description of virus detection techniques. Section 3 give summary of techniques used to evade virus detection mechanisms. Section 4 gives brief introduction of hex editing and code obfuscation techniques and their short comings. Section 3 gives detail implementation of our proposed methodology. Finally in section 4, we present our conclusion and future work.

2 Computer Virus Architecture

Generally computer virus has the following three basic building blocks [1].



```
def virus ():
    infect ()
    if trigger () is true then
       payload ()
```

Fig. 1. Pseudo code of a computer virus [1]

In Fig 1 Infect module defines how virus spreads. It selects the target to infect and defines criteria for target selection. Trigger is the condition to decide to deliver the payload or not. Payload defines the damage done by the virus. Trigger and payload are optional.

2.1 Virus Detection Techniques

This section shows some common techniques employed by virus detection tools to detect malware.

2.1.1 Signature Detection

A signature is a string of bits found in virus [3]. Signatures are found in viruses which uniquely identify them and set them apart from normal programs. Signatures are stored in signature database and antivirus tools search for these signatures in files.

2.1.2 Heuristic Analysis

New and unknown viruses can be detected using heuristic analysis techniques. It can be static or dynamic. In heuristic analysis we can analyze file format, code structure as well as we can do code emulation for virus detection. Heuristic analysis can be very noisy sometimes as it creates many false positives. Heuristic analysis is not an accurate method of detecting viruses.

Following section discusses techniques employed by malware writers to thwart detection by signature detection and heuristic analysis.

3 Code Obfuscation Techniques

To evade antivirus tools malware use different obfuscation techniques. Some are listed below.

3.1 Encryption

Encryption changes virus appearance. It consists of small decryptor stub and encrypted virus code. Virus body can be changed using different keys but decryptor remains the same, so it can be detected.

3.2 Polymorphism

To evade detection, polymorphic virus changes decryptor and virus body as well. To detect such virus code emulation techniques are used because virus body decrypts into the same virus code so it can be detected.

3.3 Metamorphism

Metamorphic viruses do not apply encryption. They change appearance of code while keeping the functionality intact. They use several code obfuscation techniques including garbage code insertion, Instruction reordering, data reordering, register renaming, subroutine in-lining, subroutine outlining, code permutation, and instruction substitution.

Commonly metamorphic viruses have embedded metamorphic engines which generate morphed copy of it using metamorphic engine. A metamorphic engine has following typical functional units as shown in Fig 2.

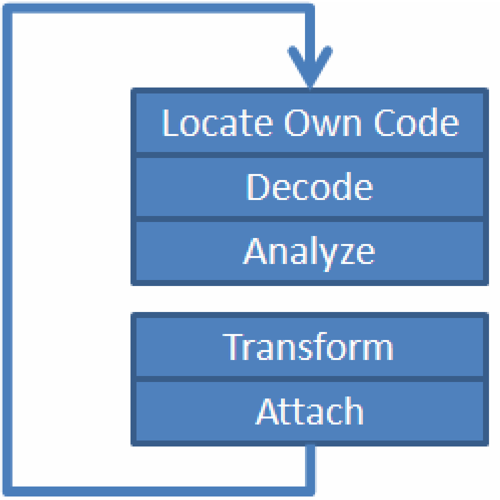
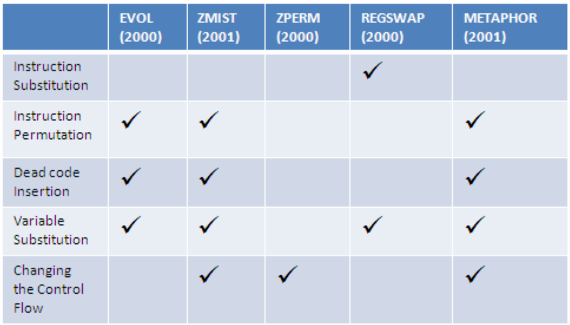


Fig. 2. Metamorphic engine functional units [4]

Metamorphic engine takes virus as input locates code to be transformed using its own customized rule set. Decode module extracts the rules by disassembling and analyze module determine transformation to be applied. Transform module applies actual transformations and attach module attaches morphed copy to a host.

Code obfuscation techniques operate on both control flow and data section of the program in assembly programs [5]. Table 1 gives summary of well known metamorphic viruses and code obfuscation techniques used by them.

Table 1. Metamorphic viruses and code obfuscation Techniques [5]



	EVOL (2000)	ZMIST (2001)	ZPERM (2000)	REGSWAP (2000)	METAPHOR (2001)
Instruction Substitution				✓	
Instruction Permutation	✓	✓			✓
Dead code Insertion	✓	✓			✓
Variable Substitution	✓	✓		✓	✓
Changing the Control Flow		✓	✓		✓

4 Hex Editing and Metamorphism

The idea behind hex editing is to find the signature in a virus code that antivirus software uses for detection and then change it [6], however, this cannot guarantee 100% virus functionality.

Commonly metamorphic viruses have embedded metamorphic engines which allow them to generate morphed copy on the fly. These metamorphic engines enable them to use extensive code obfuscation to make viruses undetectable. This technique essentially makes viruses undetectable but there is a considerable increase in virus size after each iteration [7].

5 Making It Happen: Practical Approach

We have used an amalgamation of hex editing and code obfuscation techniques to guarantee the functionality of a virus while making minimal changes to limit the size of virus in considerable bounds. Increase in size can make a virus resource-hungry. Therefore, we have used hex editing to locate the virus signature where code obfuscation needs to be applied. We have taken Evol which is a metamorphic virus itself. We will not discuss its metamorphic engine or behavior because it is out of scope of our research. The proposed technique is highly independent of Evol and can be applied on any virus to make it undetectable.

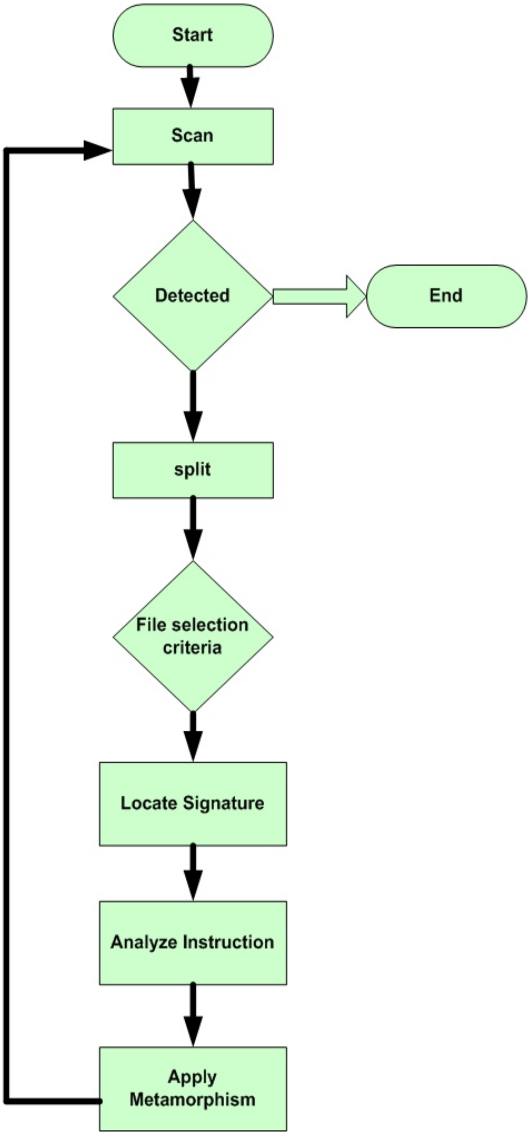


Fig. 3. Methodology Flow Diagram

5.1 Methodology

Procedure of applying metamorphic techniques is as follows:

1. Scan virus file ‘V’ which is a virus executable with an antivirus software.
2. If file ‘V’ is detected as virus
   Then

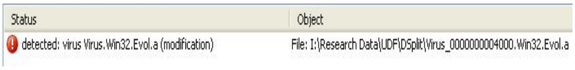
    i. Split file ‘V’ in multiple file {V1, V2… Vn}
       Such that consecutive files have one byte difference and their
       size is in incrementing order.

3. Scan files again.
4. Pick 2 files V(p-1) and Vp
   Such that p > 1
   Which satisfy these conditions
    i. V(p-1) and Vp have 1 byte difference.
    ii. Size(V(p-1)) < size(Vp)
    iii. V(p-1) is not detected by the antivirus software.
    iv. Vp is detected as a virus by the antivirus software.
5. Find location of last byte of Vp in ‘V’ and locate the assembly instruction that contains this byte.
6. Analyze instruction and apply appropriate code obfuscation technique
7. Repeat step 1 to 6 till ‘V’ is fully undetectable by the antivirus software.

Flow of our methodology is shown in Fig 3.

5.2 Implementation

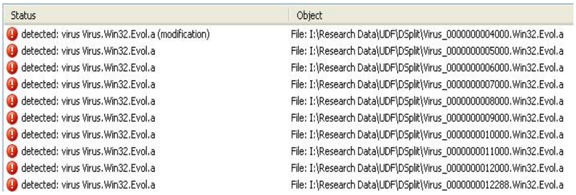
We have taken a variant of Evol virus Virus.Win32.Evol.a as a test case which is detectable by our antivirus software as shown in Fig 4.



Status	Object
detected: virus Virus.Win32.Evol.a (modification)	File: I:\Research Data\UDF\DSplit\Virus_000000004000.Win32.Evol.a

Fig. 4. Initial scan Evol detected

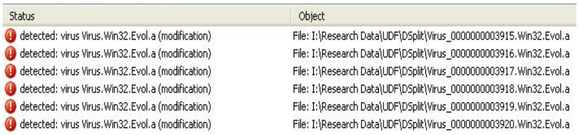
First we scanned the base virus with antivirus software and it was detected. This variant of Evol has size of 12,288 bytes. Then we split Evol viruses into files having size difference of 1000 bytes from preceding file. After splitting 13 files are generated. Each file has size 1000 bytes greater than the preceding file. Then after splitting, we scanned our files as shown in Fig 5.



Status	Object
detected: virus Virus.Win32.Evol.a (modification)	File: I:\Research Data\UDF\DSplit\Virus_000000004000.Win32.Evol.a
detected: virus Virus.Win32.Evol.a	File: I:\Research Data\UDF\DSplit\Virus_000000005000.Win32.Evol.a
detected: virus Virus.Win32.Evol.a	File: I:\Research Data\UDF\DSplit\Virus_000000006000.Win32.Evol.a
detected: virus Virus.Win32.Evol.a	File: I:\Research Data\UDF\DSplit\Virus_000000007000.Win32.Evol.a
detected: virus Virus.Win32.Evol.a	File: I:\Research Data\UDF\DSplit\Virus_000000008000.Win32.Evol.a
detected: virus Virus.Win32.Evol.a	File: I:\Research Data\UDF\DSplit\Virus_000000009000.Win32.Evol.a
detected: virus Virus.Win32.Evol.a	File: I:\Research Data\UDF\DSplit\Virus_000000010000.Win32.Evol.a
detected: virus Virus.Win32.Evol.a	File: I:\Research Data\UDF\DSplit\Virus_000000011000.Win32.Evol.a
detected: virus Virus.Win32.Evol.a	File: I:\Research Data\UDF\DSplit\Virus_000000012000.Win32.Evol.a
detected: virus Virus.Win32.Evol.a	File: I:\Research Data\UDF\DSplit\Virus_000000012288.Win32.Evol.a

Fig. 5. Scan results after splitting

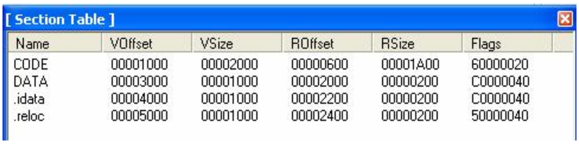
Antivirus detected all files from 4000.Win32.Evol.a to 12288.Win32.Evol.a which reveals that virus signature is certainly present in 4000.Win32.Evol.a and is not present in 3000.Win32.Evol.a. Now we repeated same process of splitting, making 3000 as start byte and 4000 as max byte having size difference of 100 bytes to narrow down the search for virus signature. After repetition of splitting and scanning process, we have narrowed down our search of signature up to file size difference of 1 byte as shown in Fig 6. The first detectable split file in incrementing order contains virus signature in the last bytes that we have to change in order to make Evol undetectable.



Status	Object
detected: virus Virus.Win32.Evol.a (modification)	File: I:\Research Data\UDF\DSplit\Virus_000000003915.Win32.Evol.a
detected: virus Virus.Win32.Evol.a (modification)	File: I:\Research Data\UDF\DSplit\Virus_000000003916.Win32.Evol.a
detected: virus Virus.Win32.Evol.a (modification)	File: I:\Research Data\UDF\DSplit\Virus_000000003917.Win32.Evol.a
detected: virus Virus.Win32.Evol.a (modification)	File: I:\Research Data\UDF\DSplit\Virus_000000003918.Win32.Evol.a
detected: virus Virus.Win32.Evol.a (modification)	File: I:\Research Data\UDF\DSplit\Virus_000000003919.Win32.Evol.a
detected: virus Virus.Win32.Evol.a (modification)	File: I:\Research Data\UDF\DSplit\Virus_000000003920.Win32.Evol.a

Fig. 6. Scan results after splitting of file size difference of 1 byte.

Now our challenge is to identify the address of signature byte and assembly language instruction which contains this byte. To locate the address of byte in executable of Evol, we can use any portable executable (PE) format editor as shown in Fig 7, whereas for assembly instruction any debugger utility can be used.



[Section Table]

Name	VOffset	VSize	ROffset	RSize	Flags
CODE	00001000	00002000	00000600	00001A00	60000020
DATA	00003000	00001000	00002000	00000200	C0000040
.idata	00004000	00001000	00002200	00000200	C0000040
.reloc	00005000	00001000	00002400	00000200	50000040

Fig. 7. Evol segments sizes

Signature byte address can be located in the executable using the following formula.

$$\alpha = \beta - \gamma + \tau + \Omega. \tag{1}$$

Where

α = RVA of Signature Byte

β = Raw Offset of Signature byte

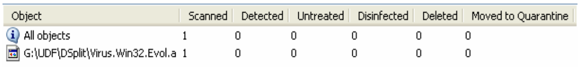
γ = Raw Offset of Section

τ = Virtual Offset of Section

Ω= Image Base

Once we have identified the location, we need to apply one of above mentioned code obfuscation technique. In our case we have applied subroutine in-lining technique to change the signature and keep the functionality of Evol unaltered.

After successful implementation of code obfuscation, our scanning results show that Evol is undetectable and fully functional as shown in Fig 8.



Object	Scanned	Detected	Untreated	Disinfected	Deleted	Moved to Quarantine
All objects	1	0	0	0	0	0
G:\UDF\DSplit\Virus.Win32.Evol.a	1	0	0	0	0	0

Fig. 8. Scan Result after applying code obfuscation

6 Conclusion

We have demonstrated successfully that malware can be made undetectable using code obfuscation techniques applying minimal changes by locating the signature. It is a challenge for antivirus community to cater this new generation of virus species that employ advanced metamorphic techniques. We have applied code obfuscation using subroutine in-lining on Evol which showed that there was no considerable increase in size. Original Evol has size of 12,288 bytes and our variant of Evol has the same size and functionality as of the original Evol virus while achieving its un-detect ability.

We proposed a methodology for producing morphed copies of a base virus that have the same functionality as the base virus and have minimal impact on size of the morphed copies. Code obfuscation is applied only where signature is detected in the base virus. Future work would be to develop a metamorphic engine that automates this process.

References

1. Aycock, J.: Computer Viruses and malware, Springer Science+Business Media (2006)
2. Cohen, F.: Computer viruses: theory and experiments. Computer Security 6(1), 22–35 (1987)
3. Stamp, M.: Information Security: Principles and Practice (August 2005)
4. Walenstein, R., Mathur, M., Chouchane, R., Lakhotia, A.: The design space of metamorphic malware. In: Proceedings of the 2nd International Conference on Information Warfare (March 2007)
5. Borello, J., Me, L.: Code Obfuscation Techniques for Metamorphic Viruses (Feburary 2008), http://www.springerlink.com/content/233883w3r2652537
6. Techotips (2009), http://techotips.blogspot.com/2009/10/tutorial-hexing-using-dsplit-hide.html
7. Desai, P.: Towards an Undetectable Computer Virus, Master's thesis, San Jose State University (December 2008)

High Flexible Sanitizing Signature Scheme Based on Bilinear Maps

Wen-Chung Kuo[1], Jiin-Chiou Cheng[2], Yen-Hung Lin[2], and Lih-Chyau Wuu[3]

[1] Department of Computer Science and Information Engineering,
National Formosa University, Yunlin 632, Taiwan, R.O.C.
simonkuo@nfu.edu.tw

[2] Department of Computer Science and Information Engineering,
Southern Taiwan University, Tainan 710, Taiwan, R.O.C.
chiou@mail.stut.edu.tw

[3] Department of Computer Science and Information Engineering,
National YunLin University of Science and Technology,
Yunlin 640, Taiwan, R.O.C

Abstract. A sanitizable signature scheme allows the sanitizer to alter the signed document using a disclosed policy that does not break the signing protocol. However, existing sanitizable signature schemes do not restrict the sanitizer, which leads to dishonest sanitizer and additional sanitizing problems. In this paper, a high flexible sanitizing signature scheme based on the bilinear mapping method that uses an arbiter to resolve the security problem is proposed. A security analysis shows that the proposed scheme retains the security requirement of sanitizable signatures and mitigates the disadvantages of related schemes.

Keywords: Sanitizable Signature, Aggregate Signature, Bilinear Maps.

1 Introduction

Almost any information can be obtained over the internet. Although current digital signature schemes, such as DSA [8], RSA [9], and ElGamal [11], can ensure the source of information and its integrity, they cannot be applied to personal privacy. When a signed document is changed, it can no longer be verified; this is known as the digital document sanitizing problem (see Fig. 1.). For example, when someone requests a patient's medical data, the hospital staff must filter sensitive information. Since the signed document was changed, the person who requested the data cannot verify that the sanitized document is the same as the original.

Ateniese *et al.* [4] proposed a sanitizable signature scheme for specific entities (called *sanitizers*) that allows partial information to be hidden in the document after it has been signed. A verifier can confirm the integrity of the disclosed parts of the sanitized document using the signature. Many other sanitizable signature schemes have been proposed [3, 4, 5, 10, 12, 15, 16]. All the schemes modify the signed document

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 402–412, 2010.

without breaking the signature agreement between the signer and verifier. The properties of the signature can be divided into disclosure condition control and non-disclosure condition control, as shown in Table 1.

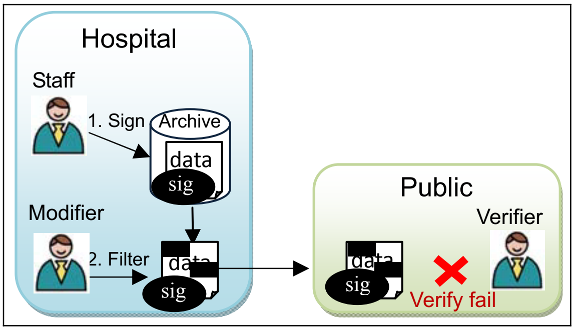


Fig. 1. Digital document sanitizing problem

Table 1. Characteristics of signatures

Property **Modification**	**Disclosure Condition Control**	**Non-Disclosure Condition Control**
Remove/Mask	MIMSYTI-scheme [6]	IKOTY-scheme [14]
Content Replace	MTS-scheme [13]	ACMT-scheme [4]

1.1 Motivation

Although the problem of confidentiality in digital signatures has been solved [8, 9, 11], new problems such as forge attacks and additional sanitizing attacks have appeared. In an additional sanitizing attack, an attacker intercepts a sanitized signed document and sanitizes portions that the sanitizer deemed undesirable. The attacker then forwards the additionally sanitized document to the verifier. Miyazaki *et al.* mentioned that a sanitizable signature scheme with disclosure condition control can avoid additional sanitizing attack [6], but the MIMSYTI scheme cannot stop a dishonest sanitizer. To solve the dispute between sanitizer and verifier, a flexible sanitizing scheme was proposed by Masubuchi *et al.* [13]. However, the flexibility of the sanitizable signature scheme is low because their disclosure condition control is set by the signer in the MTS scheme. It seems to violate the design sanitizable signature principles. For a highly flexible sanitizable signature scheme, the disclosure condition control should be set by the sanitizer.

1.2 Our Contribution

A highly flexible sanitizing signature scheme based on the bilinear maps method is proposed in this paper. The arbiter(s) resolve the dispute between the sanitizer and verifier, moreover verifier without a contract with the sanitizer. It is assumed that the sanitizer(s) can be dishonest and that the arbiter(s) is honest in the proposed scheme. A security analysis shows that the proposed scheme improves the security of the MTS scheme and solves the potential dishonest sanitizer problem presented in [7].

The rest of this paper is organized as follows: In Section 2, related works are reviewed. In Section 3, the proposed digitally signed document flexible sanitizing scheme based on bilinear maps is described. Section 4 discusses the security analysis of the proposed approach. Finally, Sections 5 the conclusion.

2 Literature Review

In this section, the MHI scheme [7] and the MTS scheme [13] are briefly introduced.

2.1 MHI-Scheme

The sanitizable signature scheme based on bilinear maps that was proposed by Miyazaki *et al.* [7] is as follows.

2.1.1 Notation

The following pairing environment is constructed:

- Construct two multiplicative cyclic groups of prime order p: G_1 and G_2, respectively.
- g_1 and g_2 are generators of G_1 and G_2, respectively.
- $\varphi(g_2)=g_1$ means φ is a computable isomorphism from G_2 to G_1.
- e is a computable bilinear map $e: G_1 \times G_2 \rightarrow G_T$ and satisfies non-degenerate.
- (y, x): the signer's public and private key which satisfies $y = g_2^x$ and $y \in G_2$.
- M: the original document.
- DID: the document ID.
- n: the number of blocks.
- $h(\cdot)$: a full-domain hash function that satisfies $\{0,1\}^* \rightarrow G_1$.
- C_i: the condition set, where i satisfies $(1 \leqq i \leqq n)$.

In addition, there are three conditions (*SANI*, *DASA,* and *DASP)* in the MHI scheme.

- *SANI* (Sanitized): portions of the signed document must be sanitized.
- *DASP* (Disclosed and Additional Sanitizing is Prohibited): portions of the signed document cannot be sanitized.
- *DASA* (Disclosed and Additional Sanitizing is Allowed): portions of the signed document can be set to condition *SANI* or *DASA*.

2.1.2 Procedure

There are three phases: signing phase, sanitizing phase, and verification phase.

- *Signing phase*:

  Step 1) Divide original document M into n subdocuments $M[i]$ $(1 \leqq i \leqq n)$.

  Step 2) Let $\tilde{M}[i] := DID \| M[i]$ and $\tilde{M}[0] := DID$ for i = 1 , …, n.

  Step 3) Compute h_i and $\sigma_i \leftarrow h_i^x$ $(0 \leqq i \leqq n)$.

  Step 4) Calculate the aggregation signature $\sigma = \prod_{i=0}^{n} \sigma_i$ $(0 \leqq i \leqq n)$ and $\sigma \in G_1$.

  Step 5) Output signature set $\{\tilde{M}[i], \sigma, \sigma_j\}$ $(0 \leqq i \leqq n$, and condition sets C_i=*DASA* $(1 \leqq i \leqq n))$.

- *Sanitizing phase*:

  Step 1) Modify the signed subdocuments based on the disclosure policy and identity of the verifier. If $M[i]$ is the sanitized subdocument, the condition is modified as follows:

  - *SANI*: Sanitizer update $\sigma = \sigma / \sigma_i$ and removes $\tilde{M}[i]$, σ_i and C_i from signature set.
  - *DASP*: Sanitizer removes σ_i and set C_i=*DASP*.

  Step 2) Output the signature set $\{\tilde{M}[i], \sigma, \sigma_j, C_j\}$ $(0 \leqq i \leqq n'$, $1 \leqq j \leqq n'$ and $n' \leqq n)$.

- *Verification phase*:

  Step 1) Compute h_i $(0 \leqq i \leqq n)$ and check whether it is identical to Eq.(1). If not, the output is invalid and the procedure terminates.

$$e(\sigma, g_2) = \prod_{i=0}^{n} e(h_i, y) \tag{1}$$

  Step 2) If a subdocument has C_i=*DASA*, it is verified using Eq.(2).

$$e(\sigma_i, g_2) = e(h_i, y) \tag{2}$$

If there is a dishonest sanitizer in the MHI scheme, additional sanitizing attacks will succeed.

2.2 MTS-Scheme

A digital document flexible sanitizing scheme was introduced by Masubuchi *et al.* (MTS scheme) [13]. Their approach mitigates the shortcomings of SUMI-4 [5] and CES (Content Extraction Signatures) [10]. A signed document which has been sanitized can be modified by a trusted third party (TTP). When the verifier disputes the sanitized document, the MTS agreement shown in Fig. 2. is used.

2.2.1 Notation

- $Flag[i]$: the condition set, where i satisfies $(1 \leqq i \leqq n)$.
- $Sig(\cdot)$: the underlying digital signature scheme.
- $Verify(\cdot)$: procedure of underlying signature verification.

The MTS scheme also uses three disclosure conditions:

- *Mandatory*: the blocks of the signed document must be sanitized (as in *SANI*).
- *Prohibited*: the blocks of the signed document cannot be sanitized (as in *DASP*).
- *Arbitrary*: the blocks of the signed document can be set to *Mandatory* or *Prohibited* (as in *DASA*).

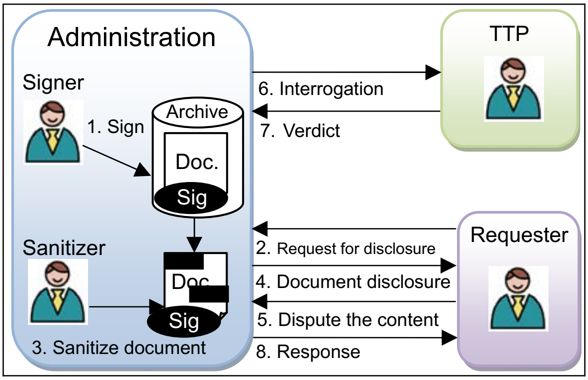


Fig. 2. MTS-scheme [13]

2.2.2 Procedure

There are four roles (signer, sanitizer, requester, and TTP(Trusted Third Party)) in the MTS scheme. The algorithm of the MTS scheme is as follows:

- *Signing phase*:

  Step 1) Divide the original document into n subdocuments m_i ($1 \leqq i \leqq n$).

  Step 2) Assign one of three conditions (*Mandatory*, *Prohibited*, or *Arbitrary*) to $flag[i]$ ($1 \leqq i \leqq n$). Then, combine all $flag[i]$ into h_f.

  Step 3) Compute $h_i = h(m_i \| r_i)$ ($1 \leqq i \leqq n$) and combine all hash values h_i into h_m.

  Step 4) Calculate $\sigma = Sig_x(Sig_x(h_m), h_f)$.

  Step 5) The signature set is $\{m_i, flag[i], \sigma\}$ ($1 \leqq i \leqq n$).

- *Sanitizing phase*:

  Step 1) Modify the signed document according to condition in $flag[i]$ ($1 \leqq i \leqq n$). The modification as follows:

  - *Mandatory*: the block m_i is replaced by h_i.
  - *Prohibited*: do nothing.
  - *Arbitrary*: chosen by sanitizer.

Step 2) Send the signature set $\{m_i', flag[i], \sigma\}$ $(1 \leqq i \leqq n)$ to the verifier.

- *Verification phase*:

  Step 1) Verify the signature set. If the signature set is available, go to step 2, otherwise, stop.

  Step 2) Verifier disputes the sanitized document and consigns the examination to the sanitizer.

- *TTP phase*:

  Step 1) Receive the dispute from sanitizer and judge its correctness. If rational, go to next step.

  Step 2) Modify $flag[i]$ and generate new h_f'.

  Step 3) Sign h_f' with its private key and send it to the sanitizer.

  Step 4) Sanitizer re-sanitizes the signed document according to $flag[i]$ that was modified by TTP and then sends it to the verifier.

Since the disclosure condition is set by the signer, the flexibility is low.

3 Proposed Scheme

This section introduces the proposed flexible sanitizable signature scheme. An arbiter is used to re-sanitize/recover the signed document. The proposed scheme supports a multi-sanitization environment and is immune to additional sanitizing attacks and the dishonest sanitizer problem.

3.1 Concept

In the proposed scheme, the three standard entities (*signers*, *sanitizer,* and *verifiers*) are considered, and a trusted third party, called an *arbiter*, is introduced. The arbiter receives a request from the verifier, re-sanitizes/recovers the signed document, and then responds to the verifier. An overview of the proposed scheme is shown in Fig. 3.

3.2 Notation

- The proposed scheme constructs the environment of pairing shown in Section 2.1.1.
- (PK_S, SK_S): Signer's public and private keys, respectively, satisfying $PK_S = g_2^{SK_S}$, $PK_S \in G_2$.
- (PK_C, SK_C), (PK_V, SK_V) and (PK_T, SK_T) are the sanitizer's, verifier's, and arbiter's keys, respectively, satisfying $PK_T = g_2^{SK_T}$.
- *En*/*De*: underlying encrypt/decrypt scheme.

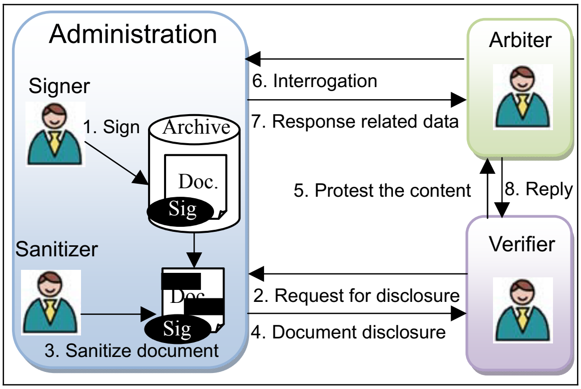


Fig. 3. Proposed scheme

There are three disclosure conditions in the proposed scheme:

- *S*: the blocks of the signed document must be sanitized (as in *SANI*).
- *P*: the blocks of the signed document cannot be sanitized (*as in DASP*).
- *A*: the blocks of signed document can be set to the condition *S* or *P* (as in *DASA*).

3.3 Procedure

In the proposed scheme, the algorithm consists of four parts: Signer (*Signing phase*), Sanitizer (*Sanitizing phase*), Verifier (*Verification phase*) and Arbiter (*Arbitration phase*). The detailed algorithm of the proposed scheme is as follows.

- *Signing phase*:

Step 1) Divide the original document m into n blocks m_i $(1 \leqq i \leqq n)$ and generate random number $\{r_1,\ldots,r_n\}$ using *seed*.

Step 2) Compute $h_i = h(m_i \| r_i)$ $(1 \leqq i \leqq n)$ and generate the individual signatures $\sigma_i = h_i^{SKs} \bmod p$ $(1 \leqq i \leqq n)$.

Step 3) Compute the aggregate signature $\sigma = \prod_{i=1}^{n} \sigma_i$.

Step 4) Let the initial condition $C_i = A$ $(1 \leqq i \leqq n)$.

Step 5) Encrypt *seed* using $En_{pkc}(seed)$.

Step 6) Output signed document $\{m_1,\ldots,m_n\}$, individual signatures σ_i $(1 \leqq i \leqq n)$, aggregate signature σ, En_{pkc} (*seed*), and condition sets C_i $(1 \leqq i \leqq n)$.

- *Sanitizing phase*:

  Step 1) If condition of m_i is *S*/*P*, the sanitizer makes the following modifications:

  - *S*: Sanitizer updates $\sigma' = \sigma / \sigma_i$ and removes m_i, σ_i, and C_i.
  - *P*: Sanitizer updates $\sigma' = \sigma / \sigma_i$ and removes σ_i and sets $C_i = P$.

  Step 2) Encrypt *seed* using $En_{PK_V}(seed)$.

  Step 3) Output sanitized document $\{m_1, \ldots, m_k\}$, individual signatures σ_i $(1 \leqq i \leqq k)$, aggregate signature σ', $En_{PK_V}(seed)$ and condition sets C_i $(1 \leqq i \leqq k)$.

- *Verification phase*:

  Step 1) Decrypt $En_{PK_V}(seed)$ by SK_V and obtain $r_1, \ldots, r_k$ using *seed.*

  Step 2) Compute $h_i = h(m_i \| r_i)$ $(1 \leqq i \leqq k)$.

  Step 3) Check whether the result is identical to Eq.(3). If yes, then accept the signature; otherwise, reject it.

$$e(\sigma', g_2) = \prod_{i=1}^{k} e(h_i, PK_S) \tag{3}$$

  Step 4) Check whether $e(\sigma_i, g_2)$ is equal to $e(h_i, PK_S)$ when the condition C_i is equal to *A*. If Eq.(4) holds, then accept the signature; otherwise, reject it.

$$e(\sigma_i, g_2) = e(h_i, PK_S) \tag{4}$$

  Step 5) When the verifier disputes the signed document, the signed document $\{m_i, C_i, \sigma'\}$ $(1 \leqq i \leqq k)$ is sent to the arbiter.

- *Arbitration phase*:

  Step 1) If the arbiter finds that the request is rational, it requests the related data $\{m_i, \sigma_i\}$ from the sanitizer.

  Step 2) Re-sanitize/recover the dispute block, if the condition of m_i is as follows:

  - *S*: disclosure block needs to be masked, arbiter updates $\sigma' = \sigma / \sigma_i$ and removes m_i, σ_i, and C_i.
  - *P*: masked block needs to disclose, arbiter computes $h_i' = h(m_i \parallel r_i')$, where r_i' is a random number. The arbiter computes $\sigma_i' = (h_i')^{SK_T}$.

  Step 3) Encrypt r_i' by using $En_{PK_V}(r_i')$.

  Step 4) Arbiter computes the signature S:$\{m_i', C_i, \sigma\}$ $(1 \leqq i \leqq k)$ or P:$\{m_i, C_i, \sigma, En(r_i'), \sigma_i'\}$ (*i* is the disputed block) for the verifier.

  Step 5) Finally, verifier verifies signature *S* using Eq.(3) or signature *P* by using Eq.(4).

The flexibility of the proposed scheme is higher than that of the MTS scheme [13] because the condition is set by the sanitizer. In addition, the verifier does not need to contact the sanitizer.

4 Security Analysis

The security of the proposed scheme was analyzed in terms of privacy, integrity, and unforgeability.

a) Privacy:

It is difficult for the verifier to recover or obtain the original document using the sanitized version.

In the sanitized signature $\{m_i', \sigma, \sigma_i, C_i, En(\text{seed})\}$, the related information m_i 、 σ_i and C_i of condition S block is removed. Therefore, only the signer and sanitizer can obtain the original document. In addition, when the arbiter wants recover the original document, it needs to contact the sanitizer. Thus, privacy is preserved in the proposed scheme.

b) Integrity:

The signature generated by the signer should be accepted by the verifier and additional sanitizing attacks should be avoided.

The verifier can compute $h_i=h(m_i \| r_i)$ and verify the signature using Eq.(3), for $i=1, \ldots, n,$ as long as the content of each block m_i $(1 \leqq i \leqq n)$ has not been modified. In other words, if an attacker forges or tampers the content m_i^* , then it will generate $h_i \neq h(m_i^* \| r_i)$. The integrity will be maintained when the sanitizer modifies the signed document because it updates $\sigma=\sigma/\sigma_i$ and removes related information using the bilinear maps' property.

c) Unforgeability:

Only the private key holder can generate a valid signature. The following two kinds of attack are considered:

1) Additional Sanitizing Attack:

An attacker wants to modify document from condition A or P to S, i.e., the attacker needs to forge σ_i corresponding to m_i and update $\sigma = \sigma / \sigma_i$. Since the signature σ_i is calculated by using h_i^{SKs}, the attacker must solve the discrete logarithm problem to get the private key SK_S. Therefore, the additional sanitizing attack is ineffective against the proposed scheme.

2) Forge Disclosure Document Attack:

An attacker wants to forge or tamper an available subdocument m_i. The attacker must generate a corresponding signature σ_i and update $\sigma=\sigma \times \sigma_i$. The attacker will encounter the problem mentioned above. Hence, the forge disclosure document attack is also ineffective against the proposed scheme.

5 Conclusion

A highly flexible sanitizing signature scheme based on bilinear maps was proposed. The approach uses an arbiter to increase flexibility. The proposed scheme is immune to both the dishonest sanitizer problem and additional sanitizing attacks.

A security analysis shows that the proposed scheme performs well in terms of privacy, integrity, and unforgeability.

Acknowledgments

This work was financially supported by the National Science Council of Taiwan under grant NSC 98-2219-E-150-001.

References

1. Brzuska, C., Fischlin, M., Freudenreich, T., Lehmann, A., Page, M., Schelbert, J., Schroder, D., Volk, F.: Security of Sanitizable Signatures Revisited. In: Jarecki, S., Tsudik, G. (eds.) Public Key Cryptography – PKC 2009. LNCS, vol. 5443, pp. 317–336. Springer, Heidelberg (2009)
2. Boneh, D., Franklon, M.: Identity based encryption from the Weil pairing. In: Kilian, J. (ed.) CRYPTO 2001. LNCS, vol. 2139, pp. 219–229. Springer, Heidelberg (2001)
3. Bonsh, D., Gentry, C., Lynn, B., Shacham, H.: Aggregate and Verifiably Encrypted Signatures from Bilinear Maps. In: Biham, E. (ed.) EUROCRYPT 2003. LNCS, vol. 2656, pp. 416–432. Springer, Heidelberg (2003)
4. Atenises, G., Chou, D.H., de Medeiros, B., Tsudik, G.: Sanitizable Signature. In: di Vimercati, S.d.C., Syverson, P.F., Gollmann, D. (eds.) ESORICS 2005. LNCS, vol. 3679, pp. 159–177. Springer, Heidelberg (2005)
5. Miyazaki, K., Susaki, S., Iwamura, M., Matsumoto, T., Sasaki, R., Yoshiura, H.: Digital Document Sanitizing Problem. IEICE Technical Report, ISEC2003-20, 61-67 (2003)
6. Miyazaki, K., Iwamura, M., Matsumoto, T., Sasaki, R., Yoshiura, H., Tezuka, S., Imai, H.: Digitally Signed Document Sanitizing Scheme with Disclosure Condition Control. IEICE Trans. Fundamentals E88-A(1), 239–246 (2005)
7. Miyazaki, K., Hanaoka, G., Imai, H.: Digitally Signed Document Sanitizing Scheme Based on Bilinear Maps. In: ASIACCS, pp. 343–354 (2006)
8. National Institute of Standards and Technology, Digital Signature Standard (DSS) (1991)
9. Rivest, R., Shamir, A., Adleman, L.: A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. Comm. of the ACM 21(2), 120–126 (1978)
10. Steinfeld, R., Bull, L., Zheng, Y.: Content Extraction Signatures. In: Kim, K.-c. (ed.) ICISC 2001. LNCS, vol. 2288, p. 285. Springer, Heidelberg (2002)
11. ElGamal, T.: A public Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms. IEEE Trans. on Information Theory IT-31(4), 469–472 (1985)
12. Izu, T., Kanaya, N., Takenaka, M., Yoshioka, T.: PIATS: A Partially Sanitizable Signature Scheme. In: Qing, S., Mao, W., López, J., Wang, G. (eds.) ICICS 2005. LNCS, vol. 3783, pp. 72–83. Springer, Heidelberg (2005)
13. Masubuchi, T., Takatsuka, M., Sasaki, R.: A Digital Document Flexible Sanitizing Scheme. In: IEEE Conference on IIH-MSP, pp. 89–92 (December 2006)

14. Izu, T., Kunihiro, N., Ohta, K., Takenaka, M., Yoshioka, T.: A Sanitizable Signature Scheme with Aggregation. In: Dawson, E., Wong, D.S. (eds.) ISPEC 2007. LNCS, vol. 4464, pp. 51–64. Springer, Heidelberg (2007)
15. Yuen, T.H., Susilo, W., Liu, J.K., Mu, Y.: Sanitizable Signature Revisited. In: Franklin, M.K., Hui, L.C.K., Wong, D.S. (eds.) CANS 2008. LNCS, vol. 5339, pp. 80–97. Springer, Heidelberg (2008)
16. Izu, T., Kunihiro, N., Ohta, K., Sano, M., Takenaka, M.: Sanitizable and Deletable Signature. In: Chung, K.-I., Sohn, K., Yung, M. (eds.) WISA 2008. LNCS, vol. 5379, pp. 130–144. Springer, Heidelberg (2009)

An Access Control Framework of Federated IPTV Providers for Mobile IPTV Services

María Elizabeth Aldana Díaz and Eui-Nam Huh

Department of Computer Engineering, Kyung Hee University,
Global Campus, South Korea
{maeliz,johnhuh}@khu.ac.kr

Abstract. M-IPTV service provision depends on different administrative domains to deliver individualized service and timely/on-demand and forces service providers to use effective mechanisms and strategies of resource management in order for them to be able to guarantee the quality levels their customers' demands during service provisioning. This paper, proposes a protocol for secure M-IPTV service provision delegation to support IPTV provider' SLAs through access control extension to different IPTV provider' security domains using Single Sign-On (SSO) technique along with handling individualized policies helping to improve communication and collaboration in B2B environment and providing users with a more consistent unified experience and seamless logon.

Keywords: Mobile IPTV, Service Level Agreements, Access control, Single Sign-on, SAML.

1 Introduction

IPTV dependencies of success have been analyzed from consumer and supplier perspective [1]. Truly interactive experience, attractive service, content-driven platform and availability of the technology to actually utilize all that it offers will guarantee IPTV success from consumer point of view. Accordingly, the ability to personalize content also gives the various service providers the ability to gather all sorts of information regarding a consumer's particular preferences increasing their use of more personalized targeted advertising, with a potentially higher chance of instigating sales.

Shin [19] study shows the user factors driving the adoption of IPTV and classifies them in intrinsic factors i.e. seeking high quality, content-rich, and value added services; and extrinsic factors i.e. highly interactive services and interoperable applications with other devices and platforms. Taking all the variables into account, the results of logistic regression shows that quality of content (special individualized service and timely/on-demand) and interactive services (value-added service and compatibility) are indeed significant predicators of the diffusion of IPTV.

IPTV services are originally targeted to fixed terminals such as set-top boxes used by different users; however, issues on the requirements for mobility support

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 413–428, 2010.

were raised as an out-growth under the auspices of the Fixed-Mobile Convergence (FMC) trend. Therefore, new personalized IPTV targets have came up to expand its value such as Mobile IPTV (M-IPTV) that enables users to transmit and receive multimedia traffic including television signal, video, audio, text and graphic services through IP-based the wired and wireless networks with support for QoS/QoE, security, mobility, and interactive functions. Through Mobile IPTV, users can enjoy IPTV services anywhere and even while on the move. There are four M-IPTV approaches which have been developed [18]: Mobile TV plus IP, IPTV plus Mobile, Cellular and Internet. Mobile TV plus IP is a convergence service between broadcasting, telecommunications and computing. IPTV plus Mobile is dominated by Telco giants in an attempt to find a new source of cash-in. Cellular is represented by Open Mobile Alliance [4] initiative defining end-to-end framework for mobile broadcast. At last, Internet approach known as Internet TV or Web TV has a short coming: that the quality of services is not guaranteed. However, considering its rapid adaptation to customer needs, this approach may be dominant in the near future. As long as mobile device uses Internet, users can access to IPTV service through various wireless access networks.

In multimedia services, security and privacy issues are urgent to be solved, such as content security and service protection. To solve these issues, some means have been proposed, such as conditional access and digital rights management. Lian's [11] work provide a digital rights management scheme for the convergent services through scalable encryption and transcoding, various business models and encryption modes and adaptive decryption.

M-IPTV service provision depends on different administrative domains to deliver individualized service and timely/on-demand and forces service providers to use effective mechanisms and strategies of resource management in order for them to be able to guarantee the quality levels their customers' demands during service provisioning. Service level agreements (SLA) are the most common mechanism used to establish agreements on the quality of a service (QoS) between a service provider and a service consumer. From the user's point of view, he expects to watch a TV program related to his preferences over his mobile phone. From IPTV provider, he makes sure the requested content is available before authorize the service, if is not, asks other IPTV provider who has same content to deliver it in order to avoid SLA violations. The key challenge for this scenario is SLAs adaptation, when IPTV provider lacks of capabilities and resources to deliver the agreed service sends consumer's SLA profile including user's policy related attributes to other IPTV provider delegating service provision to other providers efficiently. To authenticate the user Single Sign-On (SSO) [9] is a good way to access other systems. SSO is an access control of multiple, related, but independent software systems. With this property a user logs in once and gains access to all systems without being prompted to log in again at each of them.

Therefore in this work we propose to support IPTV provider's SLA through access control extension to different IPTV provider's security domains using Single sign-on technique along with handling individualized policies.

This paper is organized as follows. Section 2 provides a detailed state of the art overview of M-IPTV authentication methods, SSO approaches and SLA's implementation scenarios. In Section 3, we introduce the M-IPTV framework and secure service protocol. The performance evaluation is analyzed in Section 4. Finally, conclusions are presented in Section 5.

2 Related Work

2.1 M-IPTV Access Control

M-IPTV access control is a process by which the use of mobile multimedia resources is regulated according to a security policy and the result is a set of authorized interactions a subscriber can do with it.

Till now, efficient service protection protocols regarding IPTV by means of mobile devices have been proposed, which can be classified in three types: Data Rights Management (DRM) [11,13], IP CAS [7] and subscriber authentication technologies in AAA mechanism [17]. DRM is a technology adopted to control rights of digital content that uses Conditional Access System (CAS) [4]. CAS comprises a combination of scrambling and encryption to prevent unauthorized reception. Scrambling is the process of rendering the sound, pictures and data unintelligible. Encryption is the process of protecting the secret keys that have to be transmitted with the scrambled signal in order for the descrambler to work. Entitlement Control Messages (ECMs) are combined with a service key and the result is decrypted to produce a control word that is periodically updated. Entitlement Management Messages (EMMs) are sent over-air encrypted with a unique Master Key, pre-shared in subscriber's receiver, and carries entitlement information that includes service key and expiration date for each valid subscriber. IP CAS is a technical transplantation from broadcasting cable network to IP network with negative implications in quality and service. In quality, IP packets stream suffers a lot from disorder, delay, and jitter. In service, IP network carries different types of digital service that decrease the performance. Specialized physical client security module make CAS unsuitable for M-IPTV. Different from CAS, subscriber authentication technologies in AAA mechanism utilizes for subscriber authentication, counter-based OTP and Admissible Bilinear MAP based authorization ticket method, and even when the service channel is changed from home network to foreign network, the mobile IPTV services are consistently provided by means of authorization tickets.

2.2 Single Sign-On

The basic idea of single sign-on (SSO) is access control of multiple, related, but independent software systems. With this property a user logs in once and gains access to all systems without being prompted to log in again at each of them. The SSO service acts as the wrapper around the existing security infrastructure that exports various security features like authentication and authorization.

Different approaches for Web SSO implementation have been proposed XML-Based [8,9], One Time Password [20] and Kerberos [5,6]. XML-Based approach, provide flexibility, extensibility, and interoperability between environments to be integrated and also user authentication and access control using Security Assertion Markup Language (SAML) [15,16], a standard specification which is ratified by Organization for Advancement of Structure Information Standard (OASIS) [14]. One Time Password, it entails the user to use different password for each login to establish the communication among the applications. It eliminates the necessity of setting up the new infrastructure and also the existing system requires minimal changes to incorporate the single sign-on feature in it. It does not expose user's static and long lived password directly in the network. Kerberos SSO implementation offers the ability to prove their authenticity once in a period, and to cache the fact that authentication was successful so that subsequent authentications conducted by client software on behalf of the user need not involve the user. However, access control policies may impose other requirements that could cause users to have to re-authenticate, or provide additional proofs of authenticity. Enterprise SSO makes Kerberos credentials delegation quite difficult. However, Kerberos deployments commonly only use shared secret authentication, the protocol does support other methods of authentication and the use of Kerberos can significantly degrade a Web application's performance.

XML-Based combine with SAML has advantage over Web SSO solutions since it is a standard suitable for facilitating site access among trusted security domains after single authentication. SAML provides distributed authorization and federated identity management, and does not impose a centralized, decentralized, or federated infrastructure or solution, but instead facilitates the communication of authentication, authorization, and attribute information.

2.3 Service Level Agreement

Service level agreements (SLA) are the most common mechanism used to establish agreements on the quality of a service (QoS) between a service provider and a service consumer. SLANg [10] is a SLA specification based on XML language which integrates non-functional features (service levels) of contracts between independent parties with the functional design of a distributed component system for service provisioning horizontally or vertically. Horizontal SLAs govern the interaction between different parties providing the same kind of service whereas Vertical SLAs regulate the support parties get from their underlying infrastructure, within the service provision. SLANg defines seven different types of SLA based on a service reference model, i.e. Application, Web Service, Component, Container, Storage and Network.

Application SLAs approach is proposed in [12] as a middleware architecture for enabling Service Level Agreement (SLA)-driven clustering of QoS-aware application servers. It dynamically supports application server technologies with dynamic resource management to meet application-level QoS requirements. These requirements include timeliness, availability, and high throughput and are specified in SLAs. The middleware architecture incorporates three principal

QoS-aware middleware services: a Configuration Service, a Monitoring Service, and a Load Balancing Service to balance client requests among clustered servers, based on the actual load of those servers, thus preventing server overloading. The size of the cluster can change at runtime, in order to meet nonfunctional application requirements specified within hosting SLA. Web Service SLA [3] is intended for an enterprise server (or cluster) working as a web services provider, which supplies a collection of services through Internet to servers of other enterprises. The operational environment defines C2B connections between end clients and consumer servers, and B2B relationships between consumer servers and provider servers. Therefore, when the current load supported by the cluster is below the maximum admissible load determined from the SLA, the QoS control mechanism does not reject any new session requests. This study demonstrates that the QoS control mechanism carries out an effective differentiation of the service provided to consumers, reserving the processing capacity of the cluster for the preferential consumers during the overload periods. Moreover, the QoS control mechanism does not produce over reservation of processing capacity for the preferential consumers when the cluster operates under normal load conditions. This mechanism considers classes of requests and categories of consumers; it also guarantees the SLAs during overloads, giving priority to the service of preferential consumers.

Dynamic Networking SLAs [2] can take place between User and a Network Provider Agent which enables dynamic and flexible bandwidth reservation schemes on a per-user or per application basis. This architecture facilitates quantitative bandwidth, session duration, session start time, preferences, negotiations per user or per flow basis via SLA. The results show that these schemes can be exploited for the benefits of both negotiating parties such as getting the highest individual SLA optimization in terms of QoS and price. It is shown that in most cases, negotiation reduces rejection probability.

The fact that different types of service's SLAs are determined in XML language makes it easily extensible to increase expressiveness of non-functional features of contracts between independent parties with the functional design of a distributed component system for service provisioning.

3 Proposal

3.1 M-IPTV Framework

We design a framework based on M-IPTV service provision where mobile subscriber expects to watch a TV program related to his preferences over his mobile phone. IPTV provider makes sure the requested content is available before authorize the service, if is not, asks other IPTV provider who has same content to deliver it in order to avoid SLA violations. The key challenge for this scenario is SLAs adaptation, when IPTV provider lacks of capabilities and resources to deliver the agreed service sends consumer's SLA profile to other IPTV provider delegating service provision.

Following the above-described scenario, this framework represents an identity federation where users coming from the home IPTV provider access protected resources offered by another IPTV provider belonging to the same federation. IPTV providers belonging to identity federation demand a finer user access control in order to offer value-added services: in this case, special individualized service and timely/on-demand. M-IPTV federation scenario focuses on the protection of high-level services offering authentication mechanisms for end users, based on login/password which can be enhanced using SSO to access federation resources without further re-authentication. Another characteristic in the federation deployed is the use of access control mechanisms based on the user information i.e. age, gender and SLA which are defined in their home IPTV provider and are called user attributes.

This framework is a novelty solution which offers access control architecture to protected resources inside federation and provides mechanisms to manage both user authentication and authorization. The former is based on traditional methods and the latter on the use of authorization management techniques, making use of the user attributes, defined in their home IPTV provider to extend the federation allowing differentiated services provision to end users. SSO mechanism uses a token obtained during access, and then can be used to gain access to other services offered by the IPTV providers belonging to the federation. Figure 1 shows M-IPTV framework. The participant entities are Subscriber, Mobile Communication Network and IPTV provider federation.

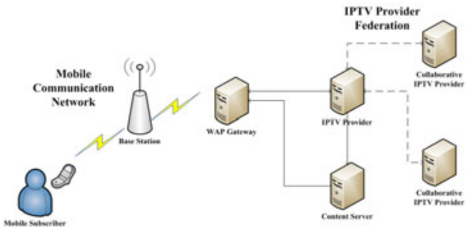


Fig. 1. M-IPTV framework

To provide access control based on user attributes, it is necessary to introduce the following features: first, IPTV providers need to define which user attributes (type and value) they are going to be responsible for that task (SLA), second, we have to define the protocol to service provision delegation. We can transport those requests through the same channel used to exchange authentication requests. Moreover, in order to provide a generic and extensible authorization environment, it would be desirable to make use of a generic framework able to hide the implementation details of the different identity management solutions deployed by each IPTV provider.

In order to provide SSO functionalities we need to cover the following issues: some kind of token needs to be defined in order to provide services with a way to be aware of the users who have been successfully authenticated and for whom no new authentication process is required. Also, the token is user transparent where the SSO process should be managed by the federation components themselves.

The following sections describe the different underlying components of the proposed framework.

3.2 Security Assertion Markup Language

SAML [15] is an XML-based framework for communicating user authentication, entitlement, and attribute information. It allows business entities to make assertions regarding the identity, attributes, and entitlements of a subject (an entity that is often a human user) to other entities, such as a partner company or another enterprise application. SAML's components [16] are:

Assertions: SAML allows for one party to assert security information in the form of statements about a subject. It contains some basic required and optional information that applies to all its statements, and usually contains a subject of the assertion (if not present, the identity determined through other means, e.g. the certificate used for subject confirmation), conditions used to validate the assertion, and assertion statements. SAML defines three kinds of statements that can be carried within an assertion: authentication, these are created by the party that successfully authenticated a user describing the particular means used to authenticate the user and the specific time at which the authentication took place; attribute, these contain specific identifying attributes about the subject; and authorization decision, these define something that the subject is entitled to do.

Bindings: SAML bindings detail exactly how the various SAML protocol messages can be carried over underlying transport protocols.

Protocols: SAML defines a number of request/response protocols that allow service providers to authenticate a principal or get assertions that meet particular criteria e.g. Artifact Resolution Protocol which provides a mechanism by which SAML protocol messages may be passed by reference using a small, fixed-length value called an artifact using one SAML binding (e.g. HTTP Redirect) while the resolution request and response take place over a synchronous binding, such as SOAP.

Profiles: Generally, a profile of SAML defines constraints and/or extensions in support of the usage of SAML for a particular application, the goal being to enhance interoperability by removing some of the flexibility inevitable in a general-use standard.

3.3 Architecture

The architecture, as shown in Figure 2, is based on secure service convergence scheme composed of User, Distribution Networks and Content Provider [11]. Among them, User sends M-IPTV service request using his mobile phone. The service request is transmitted over Mobile Communication Networks who acts as Distributor Network through WAP Gateway which connects the mobile domain and the wired Internet acting as protocol gateway to encode and decode from WAP-HTTP and vice versa respectively. IPTV provider authenticates and authorizes mobile subscriber request. If authorization process is successful Content Provider processes the multimedia content, including encoding, encryption, packaging and right issuing. Mobile Phone decrypts and descrambles content and mobile subscriber plays securely the content.

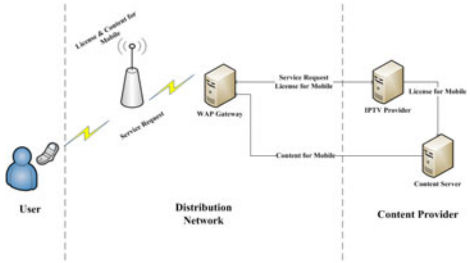


Fig. 2. Architecture of M-IPTV secure service

We propose M-IPTV architecture for secure service delegation which supports SLA adaptation. Figure 3 shows the concept of this architecture. IPTV provider receives service request, he authenticates mobile user and makes sure he counts on all resources to deliver the agreed service i.e. the requested content is available, before authorize the service. If he cannot deliver the agreed service negotiates with other IPTV provider to provide service. The negotiation is based on SSO and access control using SAML which is standardized specifications to provide flexibility, extensibility, and interoperability between environments to be integrated.

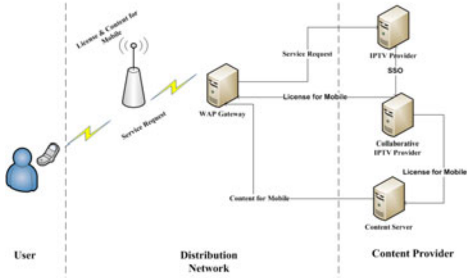


Fig. 3. M-IPTV architecture for secure service delegation

3.4 M-IPTV Secure Service Delegation Protocol

When a customer wants to use a service offered by an IPTV provider, an agreement is needed, in the same way of a traditional service. The contract involves both functional and not functional parameters relating to the service to be provided. SLA is the most common mechanism used to establish agreements on the QoS between a service provider and a service consumer.

The M-IPTV interaction process is shown in Figure 4. Firstly, the user requests to the IPTV Provider for the service. Secondly, IPTV provider authenticates the user, checks SLA to authorize the service and sends the License to the user. Thirdly, the Content Server sends content for Mobile.

The higher the number of service requests needed to be served, the higher the probability that task is not accomplished because of inability of IPTV provider to meet an SLA's objectives and the provision of a service to the customer is not successfully carried out. In such rigid context, the QoS of the final service can be strongly affected by violation on user's SLA. In order to prevent such violations SLA need to adapt during service provision with a flexible mechanism enabling the run-time negotiation of the guarantees on the QoS with other IPTV providers once violations on such guarantees are expected to occur. This would avoid both the suspension of the service provisioning and the brutal termination of the agreement.

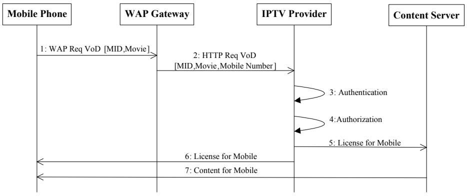


Fig. 4. M-IPTV service delivery protocol

The flexibility that we refer to consists in the possibility of (1) negotiating at run-time the service provision with SLA guarantees with others IPTV providers, and (2) accordingly delegate service delivery. This process, of course, must preserve the continuity of the service provision, i.e. the service flow must not be either interrupted or suspended while the service delegation is being negotiated.

The protocol being designed must take into account the dynamics and the new requirements that the scenario presented in this section impose. We remark that in such scenario there are several actors, in the role of IPTV provider, that stipulate one-to-one agreements with each others. We must consider:

- Mobile subscriber authentication
- M-IPTV authorization based on SLA profile
- Service delegation

Focusing on the just described requirements, in this work we add the functionality that enable the parties involved in a scenario of service delegation to negotiate SLAs guarantees while service is being provided. Figure 5 shows the concept of the proposed negotiation protocol. Firstly, the user requests to the IPTV Provider for the service. Secondly, IPTV provider authenticates the user, checks SLA to authorize the service and predicts that it does not have enough resources to deliver SLA's. Thirdly, IPTV provider redirects service request to other IPTV provider. To provide flexibility, extensibility, and interoperability between environments to be integrated, SSO based on SAML provides seamless user access to both home and collaborative IPTV security domain. IPTV collaborative provider requests a SAML authentication statement from primary IPTV provider and then, based on the authentication assertion requests an SAML attributes assertion to facilitate the SLA's terms and guarantees exchange. The terms represent contractual obligations and include a description of the service as well as the specific guarantees given that IPTV provider should assure by the availability of resources and/or on service qualities. Fourthly, collaborative IPTV provider checks SLA profile and is called to accept or reject it. If it accepts the proposal, IPTV provider will delegate service delivery; if the proposal is rejected, IPTV provider will continue asking other IPTV providers to provide M-IPTV service.

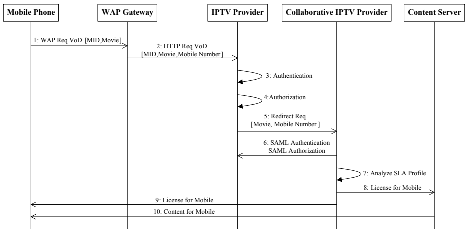

Mobile Phone
WAP Gateway
IPTV Provider
Collaborative IPTV Provider
Content Server
1: WAP Req VoD [MID,Movie]
2: HTTP Req VoD [MID,Movie,Mobile Number]
3: Authentication
4:Authorization
5: Redirect Req [Movie, Mobile Number]
6: SAML Authentication SAML Authorization
7: Analyze SLA Profile
8: License for Mobile
9: License for Mobile
10: Content for Mobile

Fig. 5. M-IPTV secure service delegation protocol

Furthermore, IPTV provider might not need to access the service offered by collaborative IPTV provider anymore (unless that service is useful to satisfy other pending requests).

4 Performance Evaluation

The proposed protocol is analyzed according to M-IPTV access control requirements and Service Level Agreement.

4.1 M-IPTV Access Control Analysis

In this section we describe and analyze the M-IPTV access control requirements. Table 1 presents the analysis result.

- Mobile user: User who can transmits and receives multimedia traffic including television signal, video, audio, text and graphic services through IP-based wireless networks with support for QoS/QoE, security, mobility, and interactive functions. The proposed architecture scheme focuses on M-IPTV service provision.

- Cross-domain authentication: Secure process which confirms user's identity with a specified level of confidence. The proposed protocol uses SSO strategy to help improve communication and collaboration in B2B environment and provides users with a more consistent unified experience and seamless logon. After successful authentication and authorization from the main logon, should be able to access external IPTV providers.
- Encryption level: It is necessary to classify the authority over access such as playing multimedia content so as to prevent any unapproved access attempt. In our design, we encrypt the NAL header only instead of doing all contents to avoid tremendous overhead on mobile devices, while DRM is not suitable for mobile devices. The detail of the mechanism is out of scope in this research.
- Multiple attributes authorization: Any information used for access control purposes. The security information described by SAML is expressed in XML format for assertions which are a declaration of a certain fact about an entity i.e. user and device. For the user, the attributes are age, gender, payment information, SLA profile, and file access permissions such as read, write and delete. Regarding the device, the attributes are type of terminal and special features.
- SLA: Service level agreements (SLA) are the most common mechanism used to establish agreements on the quality of a service (QoS) between a service provider and a service consumer. IPTV provider authorizes M-IPTV service based on SLA profile analysis to make sure it has enough resources to deliver service. If IPTV provider is unable to meet an SLA's objectives to provide service asks other IPTV provider who has same content to deliver it in order to avoid SLA violations.
- Secure service delegation: When a service request is redirected to other security domain, it considers the maintenance of privacy and identity control. IPTV provider can securely redirect M-IPTV service request in order to provide service with SLA using SSO mechanism based on SAML to provide seamless user access to both primary and collaborative IPTV security domain. Collaborative IPTV provider requests a SAML authentication statement from primary IPTV provider and then, based on the authentication assertion requests an SAML attributes assertion to facilitate the SLA's terms and guarantees exchange. Collaborative IPTV provider checks SLA profile and decides to accept or reject service provision.

4.2 Service Level Agreement

SLA adaptation process is evaluated by means of the SLA level indicator. From IPTV provider's perspective, the SLA adaptation is satisfactory if SLA's level is fair independently of the quantity of content requested. This outcome is fair if IPTV provider manages to negotiate and delegate service provision to other IPTV providers. From the customer's perspective, SLA adaptation is satisfactory if the probability that its requests are rejected or accepted with SLA violations is low and accepted requests have a high quality.

Table 1. Analysis of proposed scheme

	DRM	CAS	Proposed scheme
Mobile User	no	yes	yes
Cross-domain authentication	no	no	yes
Encryption level	heavy	heavy	light
Multiple attributes authorization	no	no	yes
SLA	not related	not related	yes
Secure service Delegation	no	no	yes

Therefore, to evaluate the satisfaction of customers and IPTV provider, the average SLA of accepted requests clearly represent fundamental performance parameters to be measured. In our model, requests increase at an average rate of three requests per second.

We compare the SLA adaptation scenario with a reference scenario without adaptation, where the IPTV provider maintains SLA's level by itself. The reference scenario represents current practice, where as providers SLA decreases and service provision stops to avoid SLA's violations. The adaptation and reference scenarios are described respectively. Figure 6 shows both scenarios where SLA's level is fair between 6 and 7 level.

SLA adaptation scenario. As discussed in Section 3, the behavior of the IPTV provider is characterized by the negotiation attitude towards collaborative IPTV providers which can be more efficient providing service depending on its available capacity and resources. The behavior of a collaborative IPTV provider is characterized by authorize or not service delegation. The negotiation is comprised of external access control and policies related multi-attributes authorization. Collaborative IPTV provider accepts service delegation and SLA's service level objectives are met to successfully carry out the provision of the service.

Reference scenario without adaptation. In the reference scenario, the mobile subscriber population places service requests at an average rate and IPTV provider verifies whether it has enough capabilities to guarantee SLA. If the capabilities are sufficient, the service request is accepted otherwise, is rejected, no adaptation is performed.

Generally, the SLA adaptation scenario exhibits a better performance than the reference scenario. In particular with higher rate of requests, that is, when capacity becomes a scarce resource, SLA is considerable fair or stable. Without adaptation, allocated capacity grows as average rate increases until saturates. The adaptation mechanism delays saturation, as capacity is allocated only to requests judged worthy by authorization process. Adaptation filters non-satisfactory capacity allocation requests only to delegate service provision fulfilling negotiation criteria. In turn, this increases the rate of accepted requests preserving the continuity of M-IPTV service provision. Otherwise, reference scenario shows that SLA

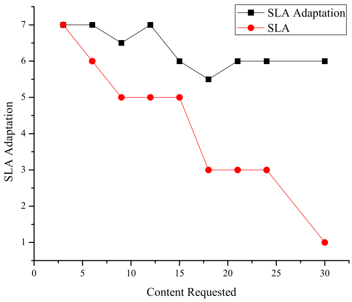


Fig. 6. Service Level Agreement scenarios

decreases proportionally to the quantity of services provided causing suspension of the service provisioning and the brutal termination of the agreement.

Adaptation provides a gain for both IPTV provider and the collaborative IPTV provider for content requested as their SLA level indicator is lower in the reference scenario. A cooperative behavior of collaborative IPTV provider allocates capacity to service requests, with positive effects when capacity is a scarce resource.

Analyzing SLA Adaptation scenario, the delivery percentage of content requested that IPTV provider has allocated at time t=1,2,3..10 at an average rate of three requests per second, ranges from 0 to 100 percentage and decreases when a new request is accepted through SLA adaptation, while it increases when a previously accepted request terminates and the associated capacity is released and used for another request. Figure 7 shows this dynamic process. In this case, 10 scenarios are presented where collaborative IPTV provider offers the same service and the same guarantees as those offered by IPTV provider. Sometimes IPTV provider might not need to access the service offered by collaborative IPTV provider, such is the case of scenario 1 where IPTV provider has enough capacity to manage 100 percentage of the content requested but when is unable to provide M-IPTV service delegates service provision to collaborative IPTV Provider and only manages 18 percentage of the content requested comparing to 88 percentage that collaborative IPTV provider manages in scenario 9.

SSO Efficiency. In the previous section we have analyzed the performance of our protocol and we have compared it with a reference scenario without adaptation. In order to complete our analysis, in this section we compare the efficiency of our access control technique based on SSO to multiple logon in case of rejection. SSO is a main component of the proposed protocol extending access control to different collaborative IPTV provider security domain without being prompted to log in again at each one of them.

SSO aims to simplify the authentication procedure. In this case, when service provision is delegated the user is not re-authenticated instead of it collaborative IPTV Provider pulls authentication and authorization information to analyze

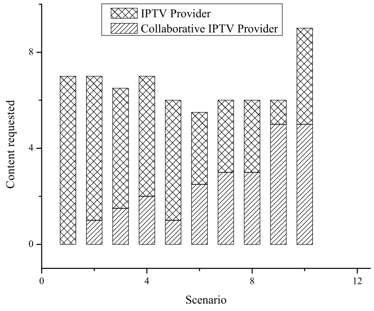


Fig. 7. SLA Adaptation service provision

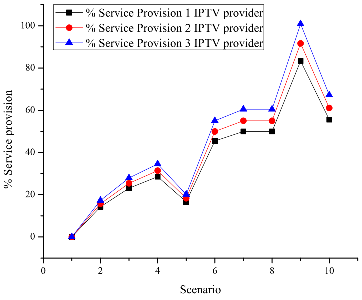


Fig. 8. Collaborative IPTV Provider percentage of service provision

whether to accept or reject service provision delegation. SSO efficiency is evaluated by means of traffic overhead i.e. the number of exchanged messages between IPTV provider and collaborative IPTV provider compare to the one caused by the combination of rejection and successfully getting the same content from alternative IPTV provider assuming mobile subscriber is a valid user for both security environments. In both scenarios we discriminate exchanged messages between entities. Results are summarized in Table 2.

Figure 8 shows the collaborative IPTV provider delivery percentage using SSO access control mechanism. It also represents the rejection percentage when SLA adaptation is not applied. The federation collaborative behavior allows dynamically balance the service provision and delivers efficiently the 100 percentage of timely-on demands.

5 Conclusions

In this work IPTV provider B2B scenario has been analyzed in order to identify access control and policies' multiple-attribute authorization requirements that M-IPTV framework needs for implementing an effective resource management mechanism for them to be able to guarantee the quality levels their customers demand during service provisioning. We have proposed the integration of new functionality to improve the flexibility of the management of SLAs in service provision. The resulting protocol exhibits a better performance in particular with higher rate of requests, that is, when capacity becomes a scarce resource increasing the rate of accepted requests and preserving the continuity of M-IPTV service provision.

Acknowledgments. This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2010-(C1090-1011-0001)) and by the IT R and D program of KCC/MKE/KEIT [Development of Open-IPTV Platform Technologies for IPTV Convergence Service and Content Sharing].

References

1. Burbridge, C.: Iptv the dependencies for success. Computer Law & Security Report 22(5), 409–412 (2006), http://www.sciencedirect.com/science/article/B6VB3-4KTPS50-9/2/5f27baff7671b6ac5b02108eb23c7e57
2. Chieng, D., Marshall, A., Parr, G.: Sla brokering and bandwidth reservation negotiation schemes for qos-aware internet. IEEE Transactions on Network and Service Management 2(1), 39–49 (2005)
3. García, D.F., García, J., Entrialgo, J., García, M., Valledor, P., García, R., Campos, A.M.: A qos control mechanism to provide service differentiation and overload protection to internet scalable servers. IEEE Trans. Serv. Comput. 2(1), 3–16 (2009)
4. Group, E.P.: Ebu project group b/ca: Functional model of a conditional access system. ebu technical review winter. Tech. rep., EBU Project Group (1995), http://www.ebu.ch/en/technical/trev/trev_266-ca.pdf
5. Group, E.P.: Ebu project group b/ca: Functional model of a conditional access system. ebu technical review winter. Tech. rep., EBU Project Group (1995), http://www.ebu.ch/en/technical/trev/trev_266-ca.pdf
6. Group, E.P.: Ebu project group b/ca: Functional model of a conditional access system. ebu technical review winter. Tech. rep., EBU Project Group (1995), http://www.ebu.ch/en/technical/trev/trev_266-ca.pdf
7. Hua, Z., Chunxiao, C., Li, Z., Shiqiang, Y., Lizhu, Z.: Content protection for iptv-current state of the art and challenges. vol. 2, pp. 1680–1685 (October 2006)
8. Jeong, J., Shin, D.: An xml-based security architecture for integrating single sign-on and rule-based access control in mobile and ubiquitous web environments. In: Meersman, R., Tari, Z., Herrero, P. (eds.) OTM 2006 Workshops. LNCS, vol. 4278, pp. 1357–1366. Springer, Heidelberg (2006), http://dx.doi.org/10.1007/11915072_39
9. Jeong, J., Shin, D., Shin, D., Oh, H.M.: A study on the xml-based single sign-on system supporting mobile and ubiquitous service environments. In: Yang, L.T., Guo, M., Gao, G.R., Jha, N.K. (eds.) EUC 2004. LNCS, vol. 3207, pp. 903–913. Springer, Heidelberg (2004), http://dx.doi.org/10.1007/978-3-540-30121-9_86

10. Lamanna, D.D., Skene, J., Emmerich, W.: Slang a language for defining service level agreements, pp. 100–106 (2003)
11. Lian, S.: Secure service convergence based on scalable media coding. Telecommunication Systems 45, 21–35 (2010), http://dx.doi.org/10.1007/s11235-009-9233-2
12. Lodi, G., Panzieri, F., Rossi, D., Turrini, E.: Sla-driven clustering of qos-aware application servers. IEEE Transactions on Software Engineering 33(3), 186–197 (2007)
13. Nishimoto, Y., Mita, N., Imaizumi, H.: Integrated digital rights management for mobile iptv using broadcasting and communications. IEEE Transactions on Broadcasting 55(2), 419–424 (2009)
14. OASIS: Organization for the advancement of structured information standards (oasis), http://www.oasis-open.org
15. OASIS: Saml v2.0 executive overview. Tech. rep., Organization for the Advancement of Structured Information StandardS (OASIS) (2005), http://www.oasis-open.org/committees/download.php/13525/sstc-saml-exec-overview-2.0-cd-01-2col.pdf
16. OASIS: Security assertion markup language (saml) v2.0 technical overview. Tech. rep., Organization for the Advancement of Structured Information Standards (OASIS) (2008), http://docs.oasis-open.org/security/saml/Post2.0/sstc-saml-tech-overview-2.0-cd-02.html
17. Park, J.: Subscriber authentication technology of aaa mechanism for?mobile iptv service offer. Telecommunication Systems 45, 37–45 (2010), http://dx.doi.org/10.1007/s11235-009-9232-3
18. Park, S., Jeong, S.H., Hwang, C.: Mobile iptv expanding the value of iptv. In: International Conference on Networking, pp. 296–301 (2008)
19. Shin, D.H.: Potential user factors driving adoption of iptv. what are customers expecting from iptv? Technological Forecasting and Social Change 74(8), 1446–1464 (2007), http://www.sciencedirect.com/science/article/B6V71-4K7WJ16-1/2/8ef3650782581658cfebd54eb7c57207
20. Tiwari, P., Joshi, S.: Single sign-on with one time password, pp. 1 –4 (November 2009)

A Novel Scheme for PMIPv6 Based Wireless Sensor Network

Md. Motaharul Islam, Sang-Ho Na, Seung-Jin Lee, and Eui-Nam Huh

Department of Computer Engineering, Kyung Hee University (Global Campus)
Youngin, Gyeonggi-do, South Korea
{motahar,davidlee}@icns.khu.ac.kr, {shna,johnhuh}@khu.ac.kr

Abstract. IP based Wireless Sensor Network (IP-WSN) is gaining tremendous importance because of its broad range of commercial applications in health care, building & home automation, asset management, environmental monitoring, security & safety and industrial automation. A network-based mobility management protocol called Proxy Mobile IPv6 has been standardized by the IETF NETLMM working group, and is starting to pay close attention among the telecommunication and Internet communities. Since host based IP mobility protocol is not feasible for the low power and low cost sensor node, network based mobility management protocol will be well suited for IP-WSN. In this paper we propose SPMIPv6 architecture, respective message formats and analyze the signaling cost and finally evaluate its performance. The result shows that the SPMIPv6 has lower signaling cost and packet delivery cost and it can improve the handover performance of UDP and TCP than the other mobility management protocol.

Keywords: NETLMM, IP-WSN, IETF, 6LoWPAN, IEEE 802.15.4.

1 Introduction

Wireless Sensor Network is comprised of a large number of sensor nodes that are densely deployed either inside the phenomenon or very close to it [1]. Advancement in the field of Wireless sensor network has enabled the development of low cost, low power, multifunctional sensor nodes that are small in size and communicate in short distances. Recently the tiny sensor nodes consisting of sensing, data processing and communicating components are capable of holding IP stack [1, 2]. That is why application of wireless sensor networks are now quite broad than the earlier. IP-WSN concept is being implemented in many sophisticated application from building and home automation to industrial manufacturing. By the introduction of adaptation layer over IEEE 802.15.4 Physical and Medium Access Control layer it becomes feasible to transmit IPv6 packet in IP-WSN [2]. Adaptation layer make usage of stateless compression technique to elide adaptation, network and transport layer header fields-compressing all the three layers down to a few bytes [3]. However IP-WSN introduces excessive signaling overhead due to its numerous tunneling over the air.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 429–438, 2010.

Excessive signaling cost becomes a barrier for the real life implementation of low power IP-WSN.

PMIPv6, a network based localized mobility management protocol provides mobility support to any IPv6 host within a restricted and topologically localized portion of the network and without requiring the host to participate in any mobility related signaling [5]. In this paper we have introduced the concept of PMIPv6 by modifying the functionality of its Mobile Access Gateway and Local Mobility Anchor to IP-WSN enabled gateway and anchor point. Then we propose the protocol architecture named SPMIPv6, its functional architecture, necessary message formats. Moreover we compare our network mobility model with 2D Random walk mobility model and finally evaluate performance of our proposed scheme.

The rest of the paper is organized as follows. Section 2 reviews the background related to PMIPv6 and 6LoWPAN. Proposed Sensor PMIPv6 Protocol architecture, sequence diagram of message flow, message formats and operational architecture are depicted in section 3. Section 4 shows performance evaluation by using an analytical model and mathematical analysis. Section 5 shows the simulation result. Finally section 6 concludes this paper.

2 Background

2.1 Overview of PMIPv6

The foundation of PMIPv6 is based on MIPv6 in the sense that it extends MIPv6 signaling and reuses many concepts such as the Home Agent (HA) functionality. However, PMIPv6 is designed to provide network-based mobility management support to a Mobile Node (MN) in a topologically localized domain. Therefore, an MN is exempt from participation in any mobility-related signaling, and the proxy mobility agent in the serving network performs mobility-related signaling on behalf of the MN. Once an MN enters its PMIPv6 domain and performs access authentication, the serving network ensures that the MN is always on its home network and can obtain its Home Address (HoA) on any access network. That is, the serving network assigns a unique home network prefix to each MN, and conceptually this prefix always follows the MN wherever it moves within a PMIPv6 domain. From the perspective of the MN, the entire PMIPv6 domain appears as its home network. Accordingly, it is not justifiable to configure the Care of Address (CoA) at the MN. The new principal functional entities of PMIPv6 are the mobile access gateway (MAG) and local mobility anchor (LMA). The MAG acts like an access router and LMA act as the mobility anchor point of the PMIPv6 domain.

2.2 6LoWPAN

6LoWPAN of IETF defines an adaptation layer for sending IPv6 packets over IEEE 802.15.4. The goal of 6LoWPAN is to reduce the size of IPv6 packets to make them fit in 127 bytes 802.15.4 frames. 6LoWPAN consists of a header compression scheme, fragmentation scheme and a method framing IPv6 Link Local Address on 802.15.4 network [2]. Also, it enhances the scalability and mobility of sensor networks. The IPv6 network defines the maximum transmission unit (MTU) as 1,280 bytes, whereas the IEEE 802.15.4 packet size is 127 octets. Therefore, the adaptation layer is defined between the IP layer and the MAC layer to transport IPv6 packets over IEEE 802.15.4 links. The adaptation layer is responsible for fragmentation, reassembly, header compression, decompression, mesh routing, and addressing for packet delivery under mesh topology. The 6LoWPAN protocol supports the scheme to compress the IPv6 header from 40 bytes to 2 bytes [8].

2.3 Problem Statement of 6LoWPAN

Devices used under 6LoWPAN are likely to be exceedingly resource constrained, and it is not desirable to enforce IP compliance directly onto the devices as is required by IP-based macro-mobility protocols; such protocols were designed to hide local mobility from networks on behalf of more powerful devices with additional resources and greater power. Continuous connectivity indicates that high signaling overhead is not appropriate for 6LoWPAN sensor nodes. In particular, if the sensor network supports mobility schemes, then excessive control signal transmission makes seamless connectivity difficult. 6LoWPAN is also unsuitable for real-time communications, making mitigation of excessive control signaling overhead an even more challenging issue. Even though the network mobility concept is suitable for 6LoWPAN mobility, as seen in the NEMO Basic Support protocol [13], the current 6LoWPAN packet format cannot support efficient mobility for a 6LoWPAN Mobile Router (MR). To support 6LoWPAN mobility, a 6LoWPAN MR needs to send a BU message and receive a BA message from it's HA; however, the 6LoWPAN packet format only defines the fragmentation and mesh routing headers. These messages are clearly not sufficient to support the mobility of a 6LoWPAN MR, since a 6LoWPAN packet does not support compressed mobility headers for BU and BA messages.

3 Proposed SPMIPv6 Protocol

3.1 Overview of SPMIPv6 Protocol

Sensor Proxy Mobile IPv6 (SPMIPv6) protocol is proposed for network based localized mobility protocol for wireless sensor network. The SPMIPv6 architecture will consists of Sensor network based Localized Mobility Anchor (SLMA), Sensor network based Mobile Access Gateway (SMAG), numerous fully functioned IPv6 header stack enabled sensor node. In our proposed SPMIPv6 protocol SLMA will also incorporate the functionality of Authentication, Authorization, and Accounting (AAA); we call it Sensor Authentication, Authorization, and Accounting (SAAA) service. The main role of SLMA is to maintain the reach ability to the sensor node's address while it moves around within the SPMIPv6 domain, and the SLMA includes a binding cache entry for each recently registered sensor node. The binding cache entry maintained at the SLMA is more specific than LMA in PMIPv6 with some additional fields such as sensor node identifier, the sensor node's home network prefix, and a flag bit indicating a sensor proxy registration. SMAG is just like an Edge Router. The main function of SMAG is to detect sensor nodes movement and initiate mobility related signaling with the sensor node's SLMA on behalf of the sensor node. Since all the devices are sensor network based so it will be energy efficient and will follow the other characteristics of 6LoWPAN. In this scheme we assume SMAG is device containing sufficient storage, processing power and unlimited power supply. The individual sensor node can either be a fully functional device containing complete IP header stack. And the other sensor node is reduced functional device. Depending on the situation the functionality of the nodes varies.

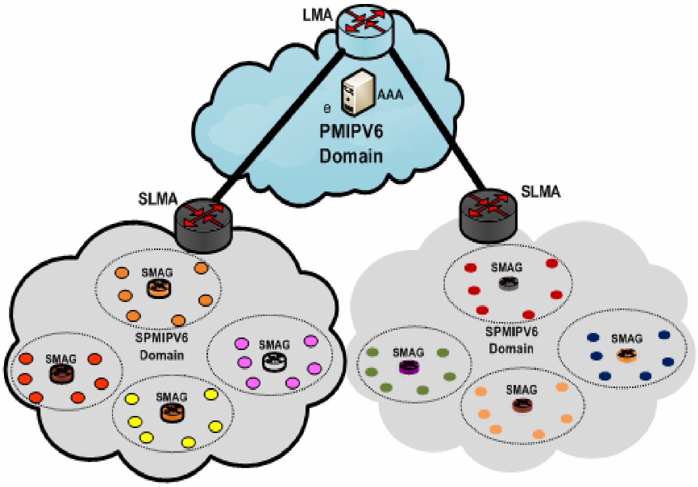


Fig. 1. Sensor Proxy Mobile IPv6 Architecture

The functionality of SLMA and SMAG in SPMIPv6 are different in many ways but similar in nature in comparison with LMA and MAG of PMIPv6. The major difference is both SLMA and SMAG works with low power 6LoWPAN sensor nodes. But both SLMA and SMAG deal with a plenty of sensor nodes. SLMA will act as a topological anchor point of all the SMAG. Inbuilt AAA functionality of SLMA helps the SMAG and sensor node to move the SPMIPv6 domain.

3.2 Message Flow in SPMIPv6

Fig 2 shows the sequence diagram of the overall messages flow in SPMIPv6. Each step shown in the sequence diagram is described as follows:

Step 1: When a sensor node first attaches to a SMAG domain, the access authentication procedure is performed using sensor node identifier via the deployed access security protocols on the access network.

Step 2 and 3: After successful access authentication, the SMAG obtains the sensor node's profile, which contains the sensor nodes ID, SLMA address, and supported address configuration mode, and so on from the policy store of SAAA service.

Step 4: Then the SMAG sends a proxy binding update (PBU) message including the MN Identifier to the sensor node's SLMA on behalf of the sensor node.

Step 5 and 6: Once the SLMA receives the PBU message, it checks the policy store to ensure that the sender is authorized to send the PBU message. If the sender is a trusted SMAG, the SLMA accepts the PBU message.

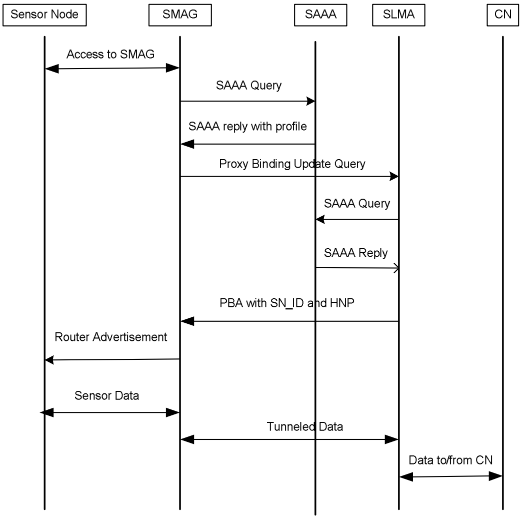


Fig. 2. Sequence diagram in SPMIPv6

Step 7: Then the SLMA sends a proxy binding acknowledgment (PBA) message including the MN’s home network prefix option, and sets up a route for the sensor node’s home network prefix over the tunnel to the SMAG.

3.3 Proxy Binding Message Format for SPMIPv6

In the proposed proxy binding update and proxy binding acknowledgement message we have added a flag bit S. If S flag is set it indicates the SMIPv6 based operations. If S bit is not set then it will indicate other operations apart from SPMIPv6.The other flags indicate meaning as mentioned in [6][7][8][9]. The mobility options field has a great significance ensuring the mobility of the sensor nodes. Depending on the scenario the mobility options field contain the respective mobility options values and facilitate the sensor mobility.

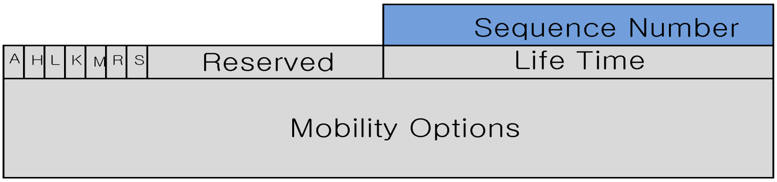


Fig. 3. SPMIPv6 PBU Message Format

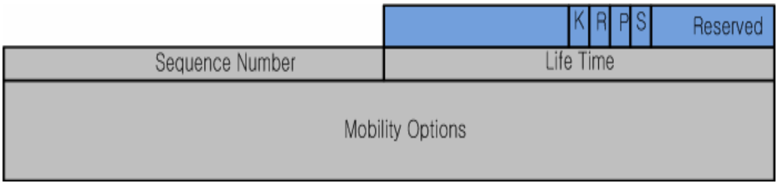


Fig. 4. SPMIPv6 PBA Message Format

3.4 Architecture of the SPMIPv6

Fig 6 represents the functional architecture of SPMIPv6 which includes the functionality of SLMA, SMAG and Sensor node. It also depicts the interaction between the three entities. Since the sensor node is IP based so consists of all the layers including adaptation layer. Sensor node will be identified by 64 bits interface identifier. And it can easily generate its IPv6 address by combining interface identifier with network prefix provided by the corresponding Sensor Mobile Access Gateway. Here SMAG is full function device that support complete implementation of IPv6 protocol stack and sensor node is reduce function device that support minimum IPv6 protocol implementation.

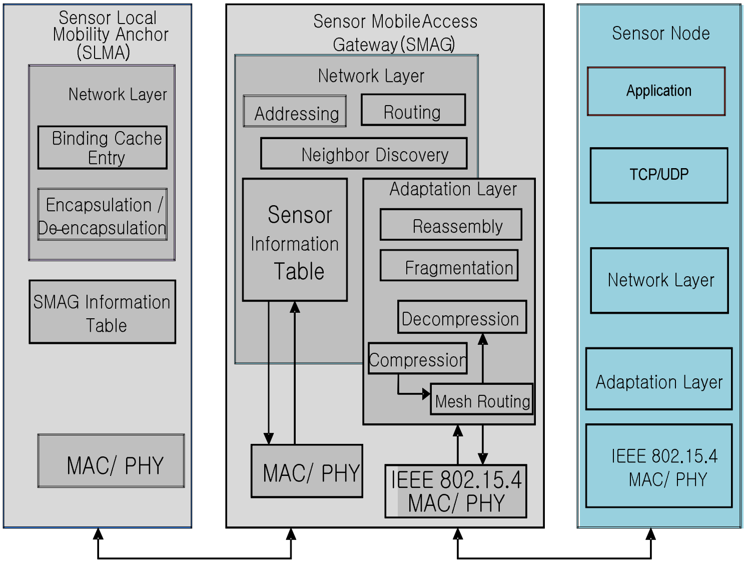


Fig. 5. Operational Architecture of SPMIPv6

4 Performance Evaluation

To evaluate the total signaling costs, we compare our analytical model with MIPv6 and SPMIPv6.

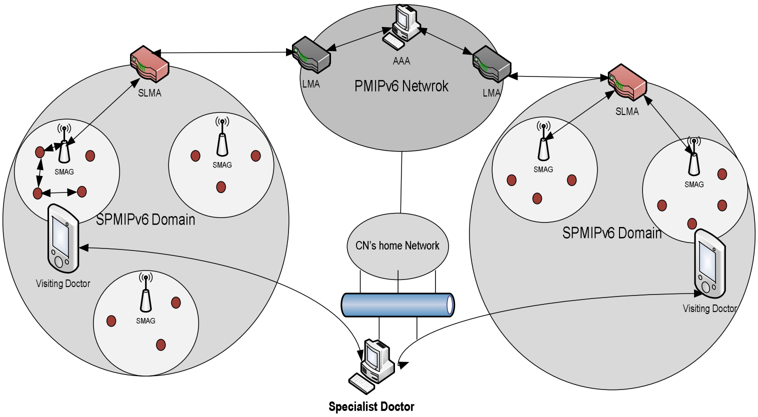


Fig. 6. Analytical model for the performances analysis of SPMIPv6

The total signaling cost (C_{mipv6}) of the proposed scheme based on MIPv6:

$$C_{mipv6} = M_{intra\_pan}.\ C_{sd} + M_{inter\_pan}.(C_{sd} + C_{bu})$$

Where

$$C_{sd} = \alpha.\ (\ RS + RA\ )\ D_{sn\text{-}smag}$$
$$C_{bu} = \beta.(BU{+}BA)\ D_{sn\text{-}smag} + \alpha.(BU{+}BA)\ D_{smag\text{-}slma}$$

The total signaling cost ($C_{spmipv6}$) of the proposed scheme based on SPMIPv6:

$$C_{spmipv6} = M_{intra\_pan}.\ C_{sd} + M_{inter\_pan}.\ (C_{sd} +\ C_{bu})$$

Where

$$C_{sd} = \alpha.\ (\ RS + RA\ )\ D_{sn\text{-}smag}$$
$$C_{bu} = \beta.\ (PBU{+}PBA)\ D_{smag\text{-}slma}$$

Table 1. System Parameter

Symbol	Description
BU	Binding Update Message
BA	Binding Acknowledgement Message
PBU	Proxy Binding Update Message

Table 1. (*continued*)

Symbol	Description
PBA	Proxy Binding Acknowledge Message
$D_{smag\text{-}slma}$	Distance between SMAG/MAG and SLMA/LMA
$D_{sn\text{-}smag}$	Distance between SN and SMAG/MAG
M_{intra_pan}	Intra PAN Mobility
M_{inter_pan}	Inter PAN Mobility
α	Unit transmission cost in wireless link
β	Unit transmission cost in wired link
RS	Router Solicitation Message
RA	Router Advertisement Message
C_{sd}	Sensor Mobility Cost
C_{bu}	Binding Update Cost

5 Simulation

In this section, we present the results of experiments evaluating the performance of our scheme, and compare the performance of our proposed scheme to MIPv6. First, we evaluated the performance of our proposed approach by mathematical analysis. Then, we set different signaling cost for the no of IP based sensor nodes in order to evaluate the consequences of our proposed scheme and MIPv6. Finally, we summarize the key characteristics of our proposed approach as compared to MIPv6 approach.

We have implemented the model and evaluate the parameter such as the signaling cost and mobility related cost presented in this paper. All of the experiments were conducted on a Windows machine with AMD Athlon(tm) 2.5 GHz CPU and 2 GB primary memory. The operating system was Microsoft Windows XP Professional Edition, and the programming tool was Visual C++ of Microsoft Visual Studio 2005.

The Fig 7 depicts the signaling cost with respect to the number of IP-WSN node in term of the MIPv6 and SPMIPv6. Signaling cost increases as the number of IP-WSN node increases. Our proposed scheme increases the performance linearly with the comparison to MIPv6. And the signaling cost increases more rapidly as the number of IP-WSN node increases. The signaling cost will be more apparent when the sensor network will be large and will cover a large geographic area.

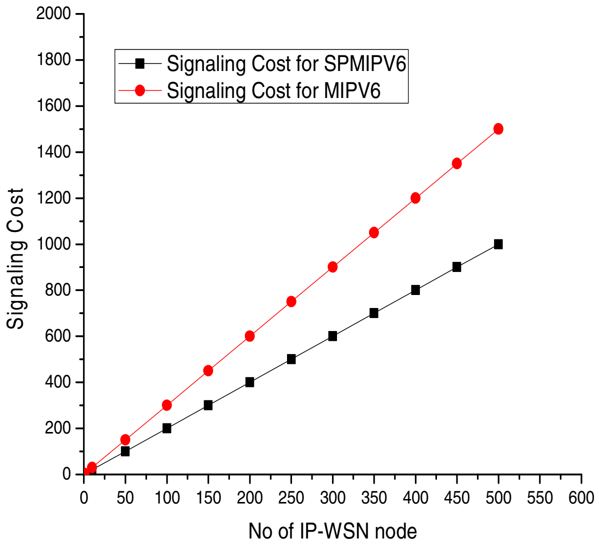


Fig. 7. Number of Node vs. Signaling Cost

6 Conclusion

Signaling cost and packet delivery for the individual tiny sensor node in IP-WSN is a big challenge to overcome. In IP-WSN, if the individual sensor node wants to communicate with the gateway router then it generates huge air traffics and it deteriorates the performance at a large scale. IETF NETLMM working group has standardized network based localized mobility management protocol called PMIPv6. In this paper we propose a network based IP-WSN scheme based on the PMIPv6 called SPMIPv6 and further develop the architecture, packet format, analyzing signaling cost and finally evaluate its performance. Analysis shows that the proposed scheme reduces the signaling cost and packet delivery cost. In this paper we only focus IP-WSN of the same vendor and protocol stack. In future we will focus on the virtualization of the sensor network consisting of multi vendor and heterogeneous protocol stack.

Acknowledgement

This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) (No. 2009-0069457) and by the MKE(Ministry of Knowledge Economy), under ITRC(Information Technology Research Center) program supervised by the IITA(Institute of Information Technology Advancement)(NIPA-2010-(C1090-1021-0003)). The corresponding author is Eui-Nam Huh.

References

1. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: A survey on sensor networks. IEEE Communications Magazine, 102–114 (2002)
2. Kushalnagar, N., Montenegro, G., Schumacher, C.: IPv6 over Low-Power Wireless Personal Area Networks (6LoWPANs): Overview, Assumptions, Problem Statement, and Goals. IETF RFC 4919 (August 2007)
3. Montenegro, G., Kushalnagar, N., Hui, J., Culler, D.: Transmission of IPv6 Packets over IEEE 802.15.4 Networks. IETF RFC 4944 (September 2007)
4. Kempf, J.: Problem statement for Network-Based Localized Mobility Management (NETLMM). IETF RFC 4830 (April 2007)
5. Johnson, D., Perkins, C., et al.: Mobility Support in IPv6, IETF RFC 3775 (June 2004)
6. Devarapalli, V., Wakikawa, R., Petrescu, A., Thubert, P.: Network Mobility (NEMO) Basic Support Protocol. IETF RFC 3963 (January 2005)
7. Soliman, H., Castelluccia, C., El Malki, K., Bellier, L.: Hierarchical Mobile IPv6 Mobility Management (HMIPv6). IETF RFC 4140 (January 2005)
8. Gundavelli, S., Leung, K., Devarapalli, V., Chowdhury, K., Patil, B.: Proxy Mobile IPv6. IETF RFC 5213 (January 2008)
9. Chalmers, R.C., Almeroth, K.C.: A Mobility Gateway for Small-Device Network. In: Hutter, D., Ullmann, M. (eds.) SPC 2005. LNCS, vol. 3450, Springer, Heidelberg (2005)
10. Akyildiz, I.F., Lin, Y.B., Lai, W.R., Chen, R.J.: A new random walk model for PCS networks. IEEE Journal on Selected Area in Communication 18(7), 1254–1259 (2000)
11. Chiang, K.H., Shenoy, N.: A 2D Random Walk Mobility Model for Location Management Studies in Wireless Network. IEEE Trans. Veh. Technol. 53(2), 413–424 (2004)
12. Kim, M.-S., Lee, S.K.: A novel load balancing scheme for PMIPv6 based networks. International Journal of Electronics and Communications (2009)
13. Shidhu, B., Singh, H.: Location Management in Cellular Networks. World Academy of Science, Engineering and Technology (2007)
14. Kim, J.H., Hong, C.S., Shon, T.: A Lightweight NEMO Protocol to Support 6LoWPAN. ETRI Journal 30(5) (October 2008)
15. Hasan, M., Akbar, A.H., Mukhtar, H., Kim, K.-H., Kim, D.-W.: A scheme to support mobility for IP based sensor networks. In: 3rd International ICST Conference on Scalable Information Systems. Vico Equense, Italy, June 4-6 (2008)
16. Shelby, Z., Bormann, C.: 6LoWPAN: The Wireless Embedded Internet. John Wiley & Sons Ltd., Chichester (2009)

Mobile Partial Identity Management: The Non-repudiable Minimal Mobile Identity Model

Mohammad Hasan Samadani and Mehdi Shajari

Computer Engineering and Information Technology Department,
Amirkabir University of Technology, Tehran, Iran
{mhssamadani,mshajari}@aut.ac.ir

Abstract. Due to the personal nature of a mobile device, it is seen as a suitable medium for storing personal information and credentials. As a result, the mobile device becomes the owner's identity-on-the-move. There are many situations in which a mobile user must offer a partial set of his identity to be able to use a service. However, only a minimal set of attributes is required by an Identity Verifier to identify a person or to correctly establish the person's role, position or status should be used. Furthermore, the verifier must be sure that the user cannot repudiate his identity later, and the offered identity is valid. In this paper, we proposed, developed and analyzed a mobile partial identity management model that the mobile user can generate his non-repudiable minimal set of his identity. The user can be sure that the verifier cannot obtain any other identity attributes, and the verifier can be sure that this identity profile is valid and the user cannot repudiate it later. The presented model, the NRMI model, is a secure and efficient mobile partial identity management. It can be used in several scenarios and use cases as well as step-by-step identification scenarios and managing identity social groups.

Keywords: Identification, Security, Privacy, Non-Repudiation, Mobile Identity, Minimal Identity, Non-Repudiable Minimal Mobile Identity Model, NRMI.

1 Introduction

Identification is essentially a process through which one answers the question "Who are you?", telling their name or other attributes. The identity of every person is composed of several personal characteristics, licenses, certificates, degrees, legal documents and business relations [1].

There are two important tools that can be used to manage the identity of users, smart cards and mobile devices. Identity-related information and credentials can be stored on smart cards and mobile devices. The security of smart cards is much higher that mobile devices. However, the management of this information can be more easily done using a mobile device. Yet another option is to securely store identity-related information on the subscriber identity module (SIM) card or another smart card attached to or embedded in the handheld device. Indeed, the smart card stores the secure information and the mobile device is used to manage this information. As a result, the mobile device becomes the owner's identity-on-the-move.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 439–449, 2010.

Ideally, only the minimal set of attributes required by an Identity Verifier to identify a person or to correctly establish the person's role, position or status should be used [1]. This requirement is called the minimal identity. Furthermore, there are many services in which the user must present his identity in a way that he cannot repudiate it later, e.g. governmental, registration, and financial services. This is the non-repudiation requirement of the identification. Also, the integrity of identity must be guaranteed by a trusted party. A safe and secure identity management system must specially satisfy these requirements.

In order to become widely accepted, a mobile identity management system must be a secure, trustworthy, usable and ubiquitous tool. Indeed, it has to work in different scenarios like proximity identification, remote identification and online transactions. Furthermore, the user must be able to easily manage his partial identities without the assistance of any other party. Indeed, the user must have the flexibility to manage his partial identity in different positions.

There are very attempts to develop a mobile identity management system. However, most of the developed systems have important shortcomings and drawbacks. In most of them, the partial identities must be created before by the identity vendor. In these models, if the user needs a new partial identity, he must contact to the identity vendor and obtain a proper partial identity profile. Also, some of them can only be used in specific scenarios, like proximity identification.

In this paper, we present a new mobile identification model which satisfies the security requirements of a safe and secure mobile identification system, e.g. privacy, non-repudiation, identity integrity, user consentience, and secure storage of sensitive information. Furthermore, this model can be easily used in different scenarios. In this model, the user has the flexibility to create a non-repudiable subset of his identity that can be easily verified by the verifier. However, none of the user and verifier must contact to identity vendor to create or be sure about the integrity and authenticity of the partial identity profile. Furthermore, the computational and communicational cost of the model is very low, which makes this model very suitable for mobile devices.

The rest of the paper is organized as follows. In the next section, we describe the mobile identity. Section 3 reviews related works. Next, in section 4 we overview the security requirements of a secure mobile identification model. Section 5 describes our presented mobile identification model and section 6 presents some use cases and scenarios. Next, in section 7 we analyze the security of the NRMI model. Section 8 discusses the implementation aspects and performance of the NRMI model. Finally, section 9 concludes the paper.

2 Mobile Device as an Identification Tool

Due to the personal nature of a hand-held device, it is seen as a suitable medium for storing personal information and credentials for access to various services. Indeed, almost all mobile phone users store their personal phone books in the memory of their devices. In addition, the devices are often used for storing personal notes, calendar items, photos, and even credit card information. As a result, the mobile device becomes the owner's identity-on-the-move, which we refer to as a *mobile identity* [1].

Identity-related information and credentials can be stored on mobile devices in different ways. This information can be easily stored in the unprotected file system of the handheld devices or in a special secure storage in them. However, the best choice could be the using of a smart card. In a mobile device, this smart card can be exactly the SIM card or a second slot one.

2.1 Terminology

The mobile identity is composed of various components which we define in this section.

Authentication is the process in which someone proves that it is who it claims to be [1].

Identification is a process in which one answers the question “Who are you?”, telling her name or other attributes. Identification is often followed by authentication, whereby the person provides a proof of her identity statement [1]. However, in most cases these two processes can be integrated into one process.

Information privacy is the claim of individuals, groups or institutions to determine for themselves when, how, and to what extent information about them can be revealed to others.

An *attribute* is a quality or characteristic of someone or something. Attributes are the building blocks of an identity. Examples of attributes are name, age, or bank account number. Biometric patterns can also be regarded as attributes [1]. Attributes have *names* (or codes) and *values*.

A *complete identity* is the union of all the person’s attributes, whereas a *partial identity* is a subset of the attributes of the complete identity [1]. Also, an *identity profile* is as a subset of user’s identity.

An *identity vendor* is an organization or company determining and attesting the attributes of an identity. Identity vendors can issue certificates attesting the connection between given attributes biometric patterns and documents. Using these certificates, persons can prove the integrity and authenticity of information about them. Note that the identity vendor is different from the identity provider. The identity provider is an organization that provides, assigns, or issues an identity or a certificate to the user, like the issuer of the National Identification Card. However, the identity vendor is an organization that attests the user’s identity and facilitates the use of this identity in electronic identification.

3 Related Works

Identity management schemes have also been developed for mobile users. An extensive study of mobile identity management schemes has been conducted by FIDIS [2]. As an example, a tool called iManager [3] enables the user to manage her partial identities on a personal digital assistant (PDA), and provides an interface for selecting the suitable identity for every use case. iManager, as most other tools, is designed for on-line business scenarios [2].

A novel architecture for mobile identity management has been presented by Hypponen, k. [1]. He presents architecture for a flexible and open mobile electronic identity tool, which can be used as a replacement for numerous ID cards and licenses. In this system the user can manage his identity profiles and request for new ones. To create a new identity profile the user must communicate with the identity vendor and request for a new profile. The most important shortcoming of this model is this communications. Indeed, for each new situation that needs a new partial identity, the user must communicate to the identity vendor.

4 Non-repudiable Minimal Identity

The most important factors in the identity management are the privacy and non-repudiation of user`s identity. To achieve the privacy and non-repudiation, the attribute set must be minimized and signed. We denote this attribute set as *non-repudiable minimal identity* of the user.

4.1 Privacy of Identity (Minimal Identity)

Ideally, only a minimal set of attributes is required by an Identity Verifier to identify a person or to correctly establish the person's role, position or status should be used. We refer to such minimal sets as *minimal identities*, and denote the requirement for using only minimal identities as *data minimization* [1].

4.2 Non-repudiation of Identity

Another important issue that must be considered, in addition to the privacy of identity, is the non-repudiation. Indeed, the user must not be able to repudiate the identity Information that he published by his consent.

There are many services in which the user must present his identity in a way that he cannot repudiate it later, e.g. governmental services, registration services, and financial services. This property can be achieved in two ways. First, the user can ask an identity vendor to guarantee his identity. In this case, the user must connect to the identity vendor and request an approved identity, i.e. a set of previously verified user`s attributes that is signed by the identity vendor. Second, the user can guarantee his identity herself. This can be achieved using user`s digital signature. This case will be discussed in next sections.

4.3 Requirements

In this section, we describe some requirements of an efficient and secure identity management system.

1. Privacy: the attribute set must be minimal.
2. Non-repudiable: the user must not be able to repudiate the presented profile.
3. User consentience: the user must be able to decide when and to who presents her partial identities. Also, he must have a full control on the profile`s attributes.
4. Computational cost: the computational cost of cryptographic operations must be low.

5. Connectivity with the identity vendor: the number of time that the user must connect the identity vendor must be low.
6. Secure storage for identity attributes: the user`s identity attributes must be stored in a secure place, e.g. SIM card.
7. Security of identity attributes: the Identity Verifier or any attacker must not be able to obtain any more attributes than the user`s consent. Indeed, the other attributes must be blinded in a way that the verifier cannot derive any more attributes.
8. Identity integrity: the user must not be able to cheat the Identity Verifier. Indeed, he must not be able to modify the complete attribute list that is verified and approved by an identity vendor.
9. Minimal modification of current systems: the system must be developed with very low modification in the available systems and models. Also, it must use the general cryptographic functions and standards, e.g. RSA, hash functions, and AES.

These requirements can be summarized into the below ones from the functional point of view:

1. The user’s information and attributes must be verified once by the identity vendor.
2. The user must be able to choose any designated subset of his attributes and make a partial profile.
3. This partial profile must be non-repudiable.
4. Making a new partial profile must be takes place locally, i.e. the identity vendor must not be involved in this process.

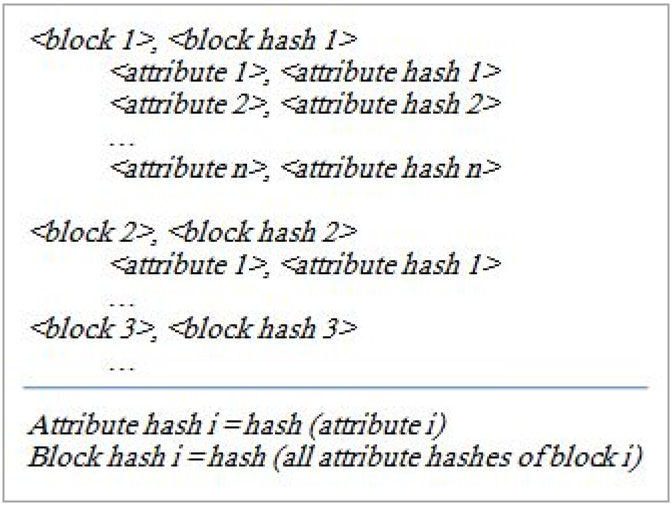
5 The Proposed Model (The Non-repudiable Minimal Mobile Identity Model)

In this section we describe our novel model in detail. The Non-Repudiable Minimal Mobile Identity model (NRMI) has three phases. In this paper, the NRMI model is presented for mobile devices. However, this model can be easily generalized to smart cards. For example, this model can be easily used in a PC with a connected smart card.

5.1 Registration to an Identity Vendor

Each user has several identity documents and information, e.g. national ID card, passport, driving certificate, location Information, health information, job Information, personal Information, and etc. In the registration phase, the user goes to an identity vendor, registers, integrates his identity information, and obtains the Complete Identity Document (CID) and Master Identity Certificate (MIC) from the identity vendor. The following describes this phase.

The user gives all necessary information, documents, and certificates to the trusted identity vendor. The identity vendor, after necessary investigations and negotiations, integrates the user’s identity information into one document, called the Complete Identity Document. We can assume this document like a XML file type. Figure 1 illustrates this document. This document consists of some blocks and each block has a number of attributes, i.e. any block groups a subset of related identity attributes. Furthermore, the identity vendor computes the hash of each attribute as the *attribute hash*, the hash of attribute hashes of each block as the *block hash*, and the hash of block hashes as the *document hash*. After that, the identity vendor creates a new document, containing the block hashes and signs it. This document is called the Master Identity Certificate. Figure 2 illustrates the MIC document.



```
<block 1>, <block hash 1>
        <attribute 1>, <attribute hash 1>
        <attribute 2>, <attribute hash 2>
        ...
        <attribute n>, <attribute hash n>

<block 2>, <block hash 2>
        <attribute 1>, <attribute hash 1>
        ...
<block 3>, <block hash 3>
        ...
```

Attribute hash i = hash (attribute i)
Block hash i = hash (all attribute hashes of block i)

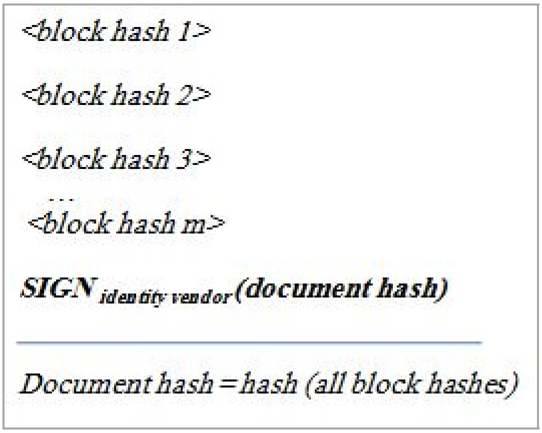
Fig. 1. The Complete Identity Document (CID)

5.2 Generation of a Non-repudiable Partial Identity

The user is able to manage his identity using a so called mobile identity management application. In a user interface, the user chooses the attributes that he wants to be shown. The identity manager makes a new document, fills it with necessary fields and attributes, and signs it. The user will have this signed document as his partial identity. The following describes this process.

The mobile identity manager application makes a new document. In this document, as illustrated in figure 3, the application puts the value of assigned attributes and the hash value of other ones. There is a new field in this document, the *assigned field*. The values for this field can be Y, N, and P. If this value is Y, it means that the related attribute or block is selected to be revealed. Otherwise, the value N means that it is not assigned. Also, if this value for a block is P, it means that some attributes of this block are not assigned and are hidden.

Finally, the application signs it. After that, the user has a Partial Identity Document (PID), and a Master Identity Certificate, issued by the identity vendor. These documents can be used for further identification processes.



<block hash 1>

<block hash 2>

<block hash 3>

. . .

<block hash m>

$SIGN_{identity\ vendor}$*(document hash)*

Document hash = hash (all block hashes)

Fig. 2. The Master Identity Certificate (MIC)

5.3 Verifying of the User Identity by the Identity Verifier

Having received the Partial Identity Document and Master Identity Certificate, the Identity Verifier is able to verify the partial identity of the user. The followings describe this process.

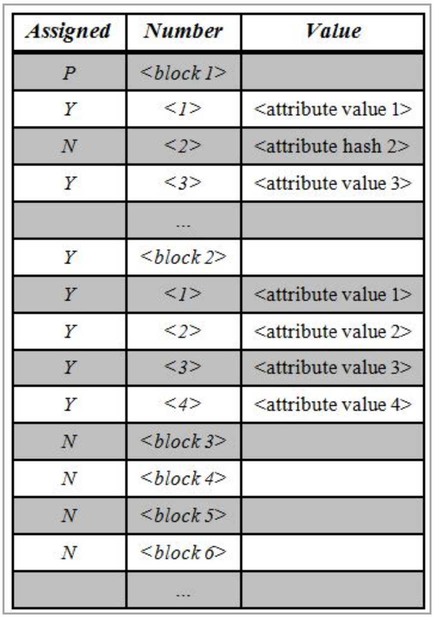
At first, the Identity Verifier checks the user's signature of the Partial Identity Document. Furthermore, the validity of the Master Identity Certificate must be checked through checking its signature. The verifier, finally, must check the authenticity of user's partial identity. To do that, the verifier traces the Partial Identity Document. For each block which the value of assigned field is N, he puts the related block hash. Also, the block hash must be computed for other remaining blocks. To compute these block hashes, the verifier computes the hash value for the attributes with known values and puts the hash value for unknown ones. Finally, the verifier computes the document hash and compares it with the document hash of Master Identity Certificate. If everything was valid, the verifier can obtain required attributes from the Partial Identity Document.

6 Use Cases

There are several use cases and scenarios in which our identification model can be applied. The following presents some examples in brief.

6.1 Secure Auction with Partially-Known Mobile Customers

Assume a secure auction in which participating have some conditions, e.g. over 18 years old, specific nationality and specific regions. However, after a user wins the auction, the server must know about his name, account number, address, phone number, postal code and other attributes. To participate in this auction, the user creates a partial identity profile with the required attributes. If the user looses the auction, he will not reveal any more attributes and the auction server and other customers only know about his age and nationality. This will provide a high level of privacy for the user. However, if the user wins the auction, he creates a new partial identity profile. This new profile will provide the necessary attributes for the auction server. Furthermore, the user reveals the least necessary information to the auction server.



Assigned	*Number*	*Value*
P	*<block 1>*	
Y	*<1>*	<attribute value 1>
N	*<2>*	<attribute hash 2>
Y	*<3>*	<attribute value 3>
	...	
Y	*<block 2>*	
Y	*<1>*	<attribute value 1>
Y	*<2>*	<attribute value 2>
Y	*<3>*	<attribute value 3>
Y	*<4>*	<attribute value 4>
N	*<block 3>*	
N	*<block 4>*	
N	*<block 5>*	
N	*<block 6>*	
	...	

Fig. 3. The identity manager puts the value of assigned attributes and the hash value of other ones into a new document, which is called PID after signing

6.2 Step-by-Step Identification

Step-by-step identification is a more general scenario than secure auction with partially-known mobile customers. Indeed, the NRMI model is very useful in scenarios in which the identity of the user must be revealed step-by-step according to events. Using the NRMI model, the user only reveals the least necessary information in each step. Therefore, if the next event not occurs, the user's privacy is guaranteed.

6.3 Mobile Friend Finder and Social Groups

Members of a social group can be divided into different sets, e.g. best friends, friends, friends of friends, unknown, invited to be friends, requested to be friends and etc. As it is obvious, each set must know different levels of information about the user. Using the NRMI model, the user can assign a partial identity profile to each set or individual. This will guarantee the privacy of the user.

6.4 Assigning of a Partial Identity to a Proxy

There are some use cases in which the user wants to delegate some of his identity attributes to his proxy. Using the NRMI model, the user can issue a Partial Identity Document to the proxy. Therefore, the proxy and the proxy environment will now only the least necessary information about the user. However, the proxy can act and be identified in behalf of the user.

The SPMS mobile signature model [4] uses the proxy certificates locally. The SIM card of the user acts as a Certificate Authority and issues a proxy certificate to the mobile device. After that, the mobile device can act as a local proxy on behalf of the user. This will provide very benefits and improvements for the mobile signature.

Integrating of the NRMI model to the SPMS model will provide some beneficial feather to the SPMS model as well as conditional full anonymity. In this case, the user and his proxy can be completely anonymous; however, they sign documents legally. If the user does some illegal acts or some forgeries, the claimant can protest against the user. In this situation the Certificate Authority and the identity vendor reveals the identity of the user. This will provide the conditional full anonymity.

As a future work, we are working on the integration of the NRMI model to our previously developed mobile signature model, the SPMS model.

7 Security Analysis and Model Evaluation

The security of the presented model must be discussed in three aspects. First, the sensitive identity information is stored in the SIM card and protected with a PIN code. Furthermore, during the partial identity creation this information never leaves the SIM card. The SIM card is a secure environment, therefore, the security of the stored data is guaranteed. Second, the created Partial Identity Documents must be sent to the verifier. However, they must not be accessible by anybody else. These documents, therefore, must be encrypted. Furthermore, the user must sign them, so the verifier can be sure about the integrity of sent data and the integrity of the origin. Third, the security of the algorithm must be guaranteed. The most important security features are that the verifier must not be able to derive any hidden identity attribute of the user from the received documents and he must be able to validate these documents.

The security of the model's algorithm is based on the security of one-way hash functions. The verifier receives the Master Identity Certificate and the Partial Identity Document from the user. Now, the verifier has these data:

1. The value of those attributes that the user wants to be revealed.
2. The hash value of those attributes that the user wants to be hidden.
3. The hash value of those blocks that the user wants to be completely hidden.
4. The signature of the identity vendor on the Document Hash.

The verifier can be sure about the validity of revealed attributes:

5. From (1) the verifier can generate the hashes of revealed attributes.
6. From (2) and (5) the verifier can generate the Block Hashed.
7. From (3) and (6) the verifier can generate the Document Hash (computed).

8. From (4) the verifier can be sure about the integrity of the Document Hash (issued by the identity vendor).
9. If these two Document Hashes, (7) and (8), are equal the verifier can be sure that the revealed data, (1), are valid and be verified by the identity vendor.

The user can be sure that the verifier cannot derive any hidden attribute from the sent documents:

10. From (2), (3), (5) and (7) the verifier knows the hashes of hidden attributes, the hashes of blocks and the hash of document. The only way to derive the value of a hidden attribute is to reverse the one-way hash function, which is impossible.

This model, furthermore, satisfy all requirements of section 4.3.

8 Implementation

We have developed a java midlet that implements the presented identification model as a proof of concept. The application has been installed on two mobile devices. The mutual connection was found using Bluetooth. The first mobile device created a partial identity and sent it to the second one. The second mobile device computed the identity of the first one.

Note that the proposed model can be implemented as an Applet in the SIM card. These two approaches vary in security and performance. If we use the SIM-based model, a higher level of security could be achieved. The performance of the cryptographic processes, however, is lower in the SIM card.

8.1 Performance Evaluation

In this section we separately analysis the performance of the three phases of the proposed method. In the first phase, the identity vendor computes the necessary hashes and generates the previously mentioned documents. The user, therefore, has no computational task to do. In the second phase, the user only selects the desired attributes in the user interface and the user's device puts their values and the hash values of the other ones in the generated document. The user, finally, signs the document. This phase, therefore, has no considerable computational cost for the user except one signature generation. In the third phase, the Identity Verifier checks some hashes and verifies some signatures. As the hash and signature verification functions are cryptographically inexpensive, this phase has very low computational cost. The results show that this phase can be done for more than 200 attributes in less than 500 milliseconds.

Another parameter that must be considered is the communication cost. The communication cost depends on the size of transferred documents. In this model, the user must only send two document to the verifier till his identity can be verified, the Master Identity Certificate and the Partial Identity Document. For a sample of 200 attributes in 10 different blocks, in which every attribute needs 100 Bytes and every hash needs 20 Bytes, the MIC is about 1 KB and the CID is about 25 KB. Also, with 50 randomly selected attributes to be revealed, the PID size is about 11 KB. As the results shows, both PID and MIC documents have a low size and can be easily transferred using Bluetooth, GPRS, and NFC. Furthermore, the CID can be easily stored in the SIM card.

9 Conclusion

The mobile identity is going to be the user's identity-on-move. Therefore, a secure and efficient mobile identification model must be developed to answer to the requirements of users. The user must be sure that the verifier cannot obtain any other identity attributes, and the verifier must be sure that this identity profile is valid and the user cannot repudiate it later. In this paper, we proposed, developed, and analyzed a new mobile partial identification model. This model satisfies the security and usability requirements of users.

Acknowledgement

This work was supported by the Iranian Telecommunication Research Center. Our thanks to Antonio Ruiz-Martínez for his valuable comments and notes.

References

1. Hyppönen, K.: An Open Mobile Identity Tool: An Architecture for Mobile Identity Management. In: Mjølsnes, S.F., Mauw, S., Katsikas, S.K. (eds.) EuroPKI 2008. LNCS, vol. 5057, pp. 207–222. Springer, Heidelberg (2008), doi:http://dx.doi.org/10.1007/978-3-540-69485-4_15
2. Müller, G., Wohlgemuth, S.: Study on mobile identity management, FIDIS (Future of Identity in the Information Society) Deliverable D3.3 (2005), http://www.fidis.net
3. Wohlgemuth, S., Jendricke, U., Gerdtom Markotten, D., Dorner, F., Müller, G.: Sicherheit und Benutzbarkeit durch Identitätsmanagement. In: Spath, D., Haasis, K. (eds.) Aktuelle Trends in der Softwareforschung - Tagungsband zum doITForschungstag, Stuttgart, pp. 241–260. IRB Verlag (2003)
4. Samadani, M.H., Shajari, M., Ahaniha, M.M.: Self-Proxy Mobile Signature, A new client based mobile signature model. In: 2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops, Perth, Australia, April 20-23 (2010)

A Privacy Preserving Service Broker Architecture for Data Sharing

Fahed Al-Neyadi and Jemal Abawajy

Deakin University,
School of Information Technology
{fmal,jemal}@deakin.edu.au

Abstract. The problem addressed in this paper is how to ensure data privacy concerns when data is shared between multiple organisations. In domains such as healthcare, there is a need to share privacy-sensitive data among autonomous but cooperating organisations. However, security concerns and compliance to privacy regulations requiring confidentiality of the data renders unrestricted access to organisational data by others undesirable. The challenge is how to guarantee privacy preservations for the owners of the information that are willing to share information with other organisations while keeping some other information secret. Therefore, there is a need for privacy preserving database operations for querying data residing at different parties. To address this challenge, we propose a new computationally efficient framework that enables organisations to share privacy-sensitive data. The proposed framework is able to answer queries without revealing any useful information to the data sources or to the third parties.

Keywords: Privacy Preservation, Data Sharing, Data Management, Privacy, Healthcare data, Database.

1 Introduction

In many domains, data integration from multiple autonomous data sources has emerged as an important practical problem [9]. For example, in order to offer the best possible care for their patients, physicians need coordinated data obtained from the physicians own patient database, from other physicians database, pharmacies, drug reference databases and labs each of which gather and maintain data for the purpose of healthcare delivery. However, security concerns and legal implications that requires privacy compliance make privacy preserving data management for querying data residing at different organisations a must. Privacy preservation means that the data owners are only willing to share the minimum information required for processing queries correctly.

In this paper, we address the problem of data sharing with privacy preserving among a diverse set of autonomous but cooperating organisations with emphases on healthcare domain. The fundamental question addressed in this paper is how can organisations share data to meet their needs to support decision making or to promote social benefits while at the same time protect personal data from being released. In the

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 450–458, 2010.

healthcare domain, the data required to support healthcare comes from several organisations such as physicians, hospitals, pharmacies and labs each of which gather and maintain data for the purpose of healthcare delivery.

Much of the data in the healthcare domain is considered private and confidential. Clearly, preservation of privacy when sharing data in a dynamic environment is very challenging. This is due to various reasons such as competition among data owners or possible legal implications requiring confidentiality of the data. There is very little work that addresses privacy concerns when data is exchanged between multiple organisations in a dynamic environment. In such environments, the approaches that enable sharing of data in a secure manner have lagged far behind the ability to store such data locally [4]. Therefore, there is a need for approaches that offer some level of control on data manipulation procedures that can respond to those privacy considerations.

In this paper, we propose a framework for data sharing with privacy preserving in a dynamic environments such as healthcare domains. The proposed framework enables the exchange of data while revealing only the minimum amount of information that is necessary to accomplish a particular task. We wish to note that the proposed framework is not domain specific, though it is perhaps most compelling for the healthcare industry because in that area there is both a desire to openly share information with anyone who needs it and a high expectation that data will not be exposed to public view or otherwise fall into the wrong hands.

The rest of the paper is organised as follows. In Section 2, we present the motivation and define the problem. We then discuss the relevant work in Section 3. In Section 4, we describe the basic idea of a privacy preserving service broker architecture. Section 5 presents the details of the proposed privacy preserving framework. We conclude with a discussion of the results and future research ideas in Section 6.

2 Motivation and Problem Statement

In this section we first motivate the need for privacy-preserving data sharing and then formulate the problem of privacy-preserving.

2.1 Motivation

In industries such as healthcare, there is a need to share privacy-sensitive data across distinct organisations. Sharing healthcare data enables early detection of disease outbreak [1], but privacy-sensitive information must be sanitised. Privacy is a complex concept and most privacy laws allow access to private information when there is adequate benefit resulting from access. For example, the Health Insurance Portability and Accountability Act (HIPPA) specify similar conditions for use of data.

Today, privacy is addressed by preventing dissemination rather than integrating privacy constraints into the data sharing process [1]. Therefore, mechanism that allows healthcare providers sharing data without revealing sensitive information about them is an important problem. For example, a healthcare provider should not release information regarding which particular patients have cancer, but it is acceptable to disclose information in regards to the existence of cancer patients in the hospital. Availability of such tools will also enable us to use distributed data in a privacy preserving way.

The importance of taking privacy into consideration when sharing personal data is demonstrated by Sweeney [6] using the medical records of Massachusetts state employees released to researchers by the Group Insurance Commission (GIC). The date was supposedly de-identified as the data did not contain identifiers such as names, social security numbers, addresses, or phone numbers and was considered safe to be made available to researchers. The data did contain demographic information such as birth date, gender, and zip code.

Unfortunately, it is not common for two individuals to have the same birth date, less common for them to also live in the same zip code, and less common still for them to also have the same gender [7]. Thus, it was easy to identify the medical history of William Weld, who was then the governor of Massachusetts, in the data provided by GIC by linking it with the voter registration list. The privacy breach occurred since GIC failed to reason about these sources. This sort of attack in which external data are linked with an anonymised data set is called a *linking attack* [6].

2.2 Problem Statement

The research challenge in developing a privacy preserving query processing solution is that the answers to the queries need to be provided while preserving the privacy of the data sources. When sharing personal information for analysis and research, healthcare providers need to ensure that individual's sensitive information should not be revealed. Specifically, the problem of query processing across multiple private databases is defined as follows [16]:

> *Given a set of autonomous data sources,* $D = \{D_1, D_2, \ldots, D_n\}$*, and a query* Q *that multiple parties collaborate to answer, the problem is to compute the answer of* Q *without revealing any additional information to any of the data sources.*

Three important properties that any privacy preserving system should provide include [4]:

a) Data privacy: The service customer learns only the answer to the query, and not any of the data used to compute it.
b) Query privacy: The service provider does not learn the query, only that a query was performed against a particular user's information.
c) Anonymous communication: Service customers and service providers do not know who the opposite party is.

Anonymous communication is important because either the query asker or the data owner might be a specialist, for example an AIDS clinic, where merely revealing that the patient is in some way associated with an organisation of that nature would constitute a privacy violation [4].

Similarly, a query might contain information about a specific condition of the patient that some of the data owners do not already know about. In this paper, we take the view that once a user has authorised an organisation to pose queries, minimum disclosure principles should be applied on a best-effort basis and not so strictly that the user might be denied proper medical care because a legitimate query could indirectly lead to a privacy violation.

In healthcare context, information about patients, disease diagnosis, medications, prevention, and treatment methods is often distributed among heterogeneous databases. The integration of these heterogeneous data sources with the objective of supporting community-wide data access is an important problem and has been addressed by a number of researchers.

In our previous work, we presented a peer-to-peer (P2P) technology-based data sharing approach that provides a decentralised data sharing infrastructure [10]. Although the approach supports fine-grained data sharing, privacy preservation was not considered in that work. In this paper, we extend our previous work such that a query processing technique over peer-to-peer systems to speed up query response time while preserving the privacy of the data owners. We exploit the fact that it is possible to perform anonymous communication in a P2P system using the overlay network.

3 Related Work

In this section, we present some of the related work in privacy. Several techniques have been proposed to preserve the privacy of data sources in the areas of databases and cryptography. For example, data perturbation is a widely known technique in data mining field to protect data privacy of statistical database [8]. Data perturbation involves adding random noise to the results of a query thus preventing the disclosure of confidential information.

A general privacy-protecting approach is to exclude all the explicit identifiers (e.g., name). An alternative approach is to use the k-anonymity[6]–and l-diversity [15] models to protect privacy. In the k-anonymity model, each record in a published table is indistinguishable from at least k-1 other records with respect to certain identifying attributes such as gender and date of birth (called quasi-identifiers). Identical quasi-identifier values constitute an equivalence class.

To achieve k-anonymity, generalisation replaces each quasi-identifier value with a more general form, which inadvertently leads to information loss. In addition, when there are many identical sensitive attribute values within an equivalence class (e.g., all patients have same disease), k-anonymity model may disclose sensitive value.

The *l-diversity* complements the *k-anonymity* by ensuring that at least l sensitive values are well-represented in every equivalence class. The problem with the k-anonymity model is that an attacker can discover the values of sensitive attributes when there is little diversity in those sensitive attributes [7]. Also, when the attackers have the background knowledge, the *k-anonymity* model does not guarantee privacy against attackers using background knowledge [7]. However, *l-diversity* may be difficult and unnecessary to achieve and does not consider semantic meanings of sensitive values.

Sharing private information across distributed databases with focus on the query execution alone is discussed in [4]. A mechanism for privacy preserving query processing using third parties is discussed in [11, 16]. Rezgui et al. [11] proposed a privacy mediator based on the screening of external queries for sensitive attributes and their eventual removal from processed queries. A class of security mediators that go beyond the state of the art with regard to the privacy protection methods applied to the final result set. We propose to extend the common "query rewriting" approach by integrating more sophisticated methods from the areas of record anonymisation and statistical disclosure control, specifically for the detection and limitation of interval disclosure.

Similar approach by the extent to which the trusted third party is used in maintaining privacy., the idea is to perturb the answer to every query [12, 13], or to precisely answer some queries; but deny the answer to others so as to ensure privacy. In the trusted third parties approach, the data owners hand over their data to a trusted third party and the third party computes the results of the queries [11]. However, this level of trust is unacceptable for privacy preserving query processing [16]. Our approach differs in that it does not hand-in data to the trusted third parties.

A mechanism to process intersection queries using oblivious transfer and polynomial evaluation based on cryptographic techniques. The scheme ensures that data sources will only know the intersection but nothing else [3]. Hacigumus et al. [17], Hore et al. [18] and Aggarwal et al. [2] propose using third parties as database service providerswithout the need for expensive cryptographic operations. However the proposed schemes do not allow queries to execute over the data of multiple providers, which is the main focus of our work.

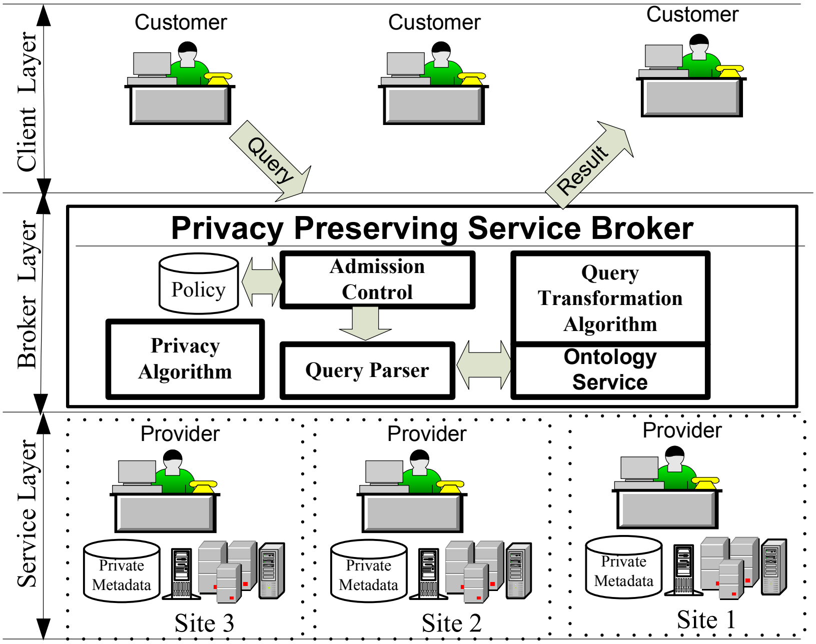


Fig. 1. High-level architecture of the privacy preserving scheme

It must be noted that our model of data privacy is one of soft constraints. We believe there will always exists a need for the healthcare providers to selectively share data from their internal records with affiliated organisations, as permitted by law and required by business processes, without requiring the patient's involvement [4]. Our work complements this work as we address both the query execution as well as the decision of determining which queries should be permitted for execution.

4 Layered System Architecture

Fig. 1 shows the high-level architecture of the system model used in this paper. The system is composed of n sites, $S = \{S_1, S_2, \ldots, S_n\}$, each site represent an autonomous healthcare organisations. Each site, $S_i \in S$, stores its own patient information that includes demographic data, diagnostic data, prescription data, treatment data, financial and insurance data. We assume that patient data is distributed throughout the system. Throughout this paper, we refer to the data requesters as *service customers* while to data providers as *service providers*.

4.1 Service Provider

Service providers are responsible for ensuring the availability and persistence of any data that they control. A service provider maintains information about its customers in some internal database with a proprietary schema. As in [4], the cases of erroneous identifiers or persons with multiple aliases are outside the scope of this paper. Also, we do not consider the use of known identifying information such as a social security number or a combination of name and age, to be a privacy threat in this context because it is typically already known by the types of organisations targeted by our work. Each service provider maintains a private metadata, which is a set of privacy views. Each privacy view defines a set of private attributes, its owner, the tables, and conditional expression.

Providers have public keys **K** and private keys **k** respectively. A provider **A** communicating with a provider **B** is able to sign messages it sends and verify the signature on messages that it receives using the following functions;

Sign (kA, msg)
Verify (KB, msg)

It may also encrypt and decrypt messages using functions

Encrypt (KB, msg)
Decrypt (kA, msg)

Additionally, providers must be able to verify that any other providers with which they communicate are authorised to participate in the system. This is accomplished by having the public keys signed by a trusted accreditation agency.

4.2 Service Consumer

Service customers run queries against data that is distributed throughout the system. Queries are written in a relational algebraic language similar to SQL. Standard select, project, and join operations are performed against tables that are fragmented across various providers and the results may range from yes/no answers computed over the data to actual values of data found in the databases.

Results must not contain any information that the customer is not authorised to discover. In this paper we focus on the query execution, leaving the decision of determining which queries should be permitted for future work. All aggregation of information across table fragments at different organizations is performed at query time. Records are kept separate in the underlying databases and only the links between them are stored; this ensures that any erroneous linkages may be undone without leaving the original data in a polluted state.

4.3 Service Broker

In this paper, we assume that each organisation is responsible for storing and maintaining the data that it generated. Customers interested in certain data submit queries to the broker. The broker is assumed to be secure and fully trusted to safeguard the privacy of organisational data. The service broker is responsible for both the query execution as well as the decision of determining which queries should be permitted for execution. The broker ensures that results of the query do not contain any information that the asker is not authorized to discover. The broker uses an algorithm called semantic request mediation based on global ontology [14] to translate customer queries so that the semantic heterogeneity between organisation A's schema and N's schema can be resolved.

5 Privacy-Preserving Data Sharing Framework

In this section, we demonstrate how queries can be answered in a distributed manner while preserving the privacy of the original data.

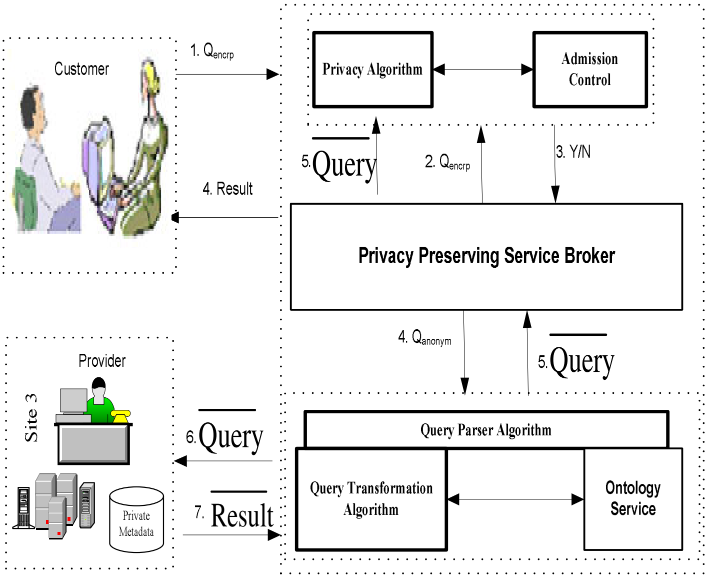

Customer
1. Q_{encrp}
Privacy Algorithm
Admission Control
5. $\overline{\text{Query}}$
2. Q_{encrp}
3. Y/N
4. Result
Privacy Preserving Service Broker
4. Q_{anonym}
5. $\overline{\text{Query}}$
Provider
Site 3
6. $\overline{\text{Query}}$
7. $\overline{\text{Result}}$
Private Metadata
Query Parser Algorithm
Query Transformation Algorithm
Ontology Service

Fig. 2. Privacy preserving broker algorithm protocol

The main idea of the proposed framework is to support queries that reveal enough information so that organisations can go about their business, and no more. For the sake of clarity and to permit better optimisation later, we split querying into two phases. Phase 1 performs a global search for records pertaining to the person in question and returns a set of data handles, each of which indicates the presence of a record somewhere in the system but does not reveal where that record resides. Phase 2 uses the data handles to execute a relational algebraic query, while keeping the original data hidden from the customer and keeping the query hidden from the data owners.

Fig. 2 shows the privacy preserving algorithm protocol in action. Potential customers encrypt the query and send it to the broker. Service customers are also required to submit their credentials to the broker who then checks (this done by the admission control algorithm) to see if the customer is authorised to post query to any of the service provider registered with the broker.

Queries that are admitted will be sent to the Query parser algorithm, which checks, among other things, if the query contains attributes that the customer is not authorised to access. If so, the parser rewrites the query and deletes all private attributes before clearing the query for execution. Queries that are cleared will then be sent to the privacy algorithm for anonmising it before sending it to service provider for processing. When the query from customer or the results of the query from the service provider arrives, the broker invokes the privacy algorithm. Also, the privacy algorithm is responsible for taking the result as input and anonmise it. This makes the broker able to answer queries without revealing any useful information to the data sources or to the third parties.

6 Conclusions and Future Work

In this paper, we addressed the problem of developing a privacy-preserving query processing solution for data sharing in a dynamic environment. We presented a broker-based solution that provides answers to the legitimate queries while preserving the privacy of the data sources. More work is needed on the details, especially on the issue of deciding which queries are allowable and which are not, but the basic technology now exists to build an infrastructure that permits sharing of sensitive data across organisation boundaries. We are currently working on several directions. First, besides data integration, the broker should also be able to enforce the privacy policies that the service providers have specified. Second, we are currently studying the proposed framework using simulations. Due to the sensitive nature of real world data, we generated our own, consisting of 40 providers and around 30, 000 patients. Published statistics [8, 9] indicate a rough power law distribution in the size of providers and in the frequency of patient visits, so the topology took this into account.

References

1. Tsui, F.-C., Espino, J.U., Dato, V.M., Gesteland, P.H., Hutman, J., Wagner, M.M.: Technical description of RODS: A real-time public health surveillance system. J. Am. Med. Inform. Assoc. 10(5), 399–408 (2003)

2. Aggarwal, G., Bawa, M., Ganesan, P., Garcia-Molina, H., Kenthapadi, K., Motwani, R., Srivastava, U., Thomas, D., Xu, Y.: Two can keep a secret: A distributed architecture for secure database services. In: CIDR, pp. 186–199 (2005)
3. Naor, M., Pinkas, B.: Oblivious transfer and polynomial evaluation. In: Proc. of the thirty-first annual ACM symposium on Theory of computing, pp. 245–254. ACM Press, New York (1999)
4. Siegenthaler, M., Birman, K.: Sharing Private Information Across Distributed Databases. In: Eighth IEEE International Symposium on Network Computing and Applications, pp. 82–89 (2009)
5. LeFevre, K., Agrawal, R., Ercegovac, V., Ramakrishnan, R., Xu, Y., DeWitt, D.: Limiting disclosure in hippocratic databases. In: VLDB 2004: Proceedings of the Thirtieth international conference on Very large data bases. VLDB Endowment, pp. 108–119 (2004)
6. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 557–570 (2002)
7. Chen, B., Kifer, D., Lefevre, K., Machanavajjhala, A.: Privacy-Preserving Data Publishing (Survey). Foundations and Trends in Databases, vol. 2, pp. 1–167 (2009)
8. Agrawal, R., Srikant, R.: Privacy-Preserving Data Mining. In: Proceedings of SIGMOD (2000)
9. Hu, N.: Privacy-Preserving Query Checking in Query Middleware, fskd, vol. In: 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, vol. 1, pp. 590–594 (2009)
10. Al-Nayadi, F., Abawajy, J.H., Deris, M.M.: A Conceptual Framework for Ubiquitously Sharing Heterogeneous Patient Information among Autonomous Healthcare Providers. In: International Conference on Multimedia and Ubiquitous Engineering (MUE 2007), pp. 299–306 (2007)
11. Rezgui, A., Ouzzani, M., Bouguettaya, A., Medjahed, B.: Preserving Privacy in WebServices. In: Proceedings of the Workshop on Web Information and Data Management (WIDM 2002), pp. 56–62 (2002)
12. Blum, A., Dwork, C., McSherry, F., Nissim, K.: Practical privacy: the sulq framework. In: PODS, pp. 128–138 (2005)
13. Dwork, C., Nissim, K.: Privacy-preserving Data Mining on Vertically Partitioned Databases. In: Franklin, M. (ed.) CRYPTO 2004. LNCS, vol. 3152, pp. 528–544. Springer, Heidelberg (2004)
14. Mitra, P., Pan, C.-C., Liu, P., Atluri, V.: Privacy preserving semantic interoperation and access control of heterogeneous databases. In: Proc. ACM Conf. on Computer and Communications Security, pp. 66–77 (2006)
15. Machanavajjhala, A., Gehrke, J., Kifer, D.: l-Diversity: Privacy beyond k-anonymity. In: International Conference on Data Engineering (ICDE), pp. 24–24 (2006)
16. Emekci, F., Agrawal, D., Abbadi, A.E., Gülbeden, A.: Privacy Preserving Query Processing using Third Parties. In: Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006 (2006)
17. Hacigumus, H., Iyer, B.R., Li, C., Mehrotra, S.: Executing SQL over encrypted data in the database service provider model. In: SIGMOD Conference (2002)
18. Hore, B., Mehrotra, S., Tsudik, G.: A privacypreserving index for range queries. In: Proc. of the 30th Int'l Conference on Very Large Databases VLDB, pp. 720–731 (2004)

Checking the Paths to Identify Mutant Application on Embedded Systems

Ahmadou Al Khary Séré[1], Julien Iguchi-Cartigny[2], and Jean-Louis Lanet[2]

[1] XLIM Labs, Université de Limoges, JIDE, 83 rue Isle, Limoges, France
ahmadou-al-khary.sere@xlim.fr
[2] Université de Limoges, JIDE, 83 rue Isle, Limoges, France
{julien.cartigny,jean-louis.lanet}@unilim.fr

Abstract. The resistance of Java Card against attack is based on software and hardware countermeasures, and the ability of the Java platform to check the correct behaviour of Java code (by using bytecode verification for instance). Recently, the idea to combine logical attacks with a physical attack in order to bypass bytecode verification has emerged. For instance, correct and legitimate Java Card applications can be dynamically modified on-card using laser beam. Such applications become mutant applications, with a different control flow from the original expected behaviour. This internal change could lead to bypass control and protection and thus offer illegal access to secret data and operation inside the chip. This paper presents an evaluation of the ability of an application to become mutant and a new countermeasure based on the runtime check of the application control flow to detect the deviant mutations. . . .

Keywords: Smart Card, Java Card, Fault Attack, Control Flow Graph.

1 Introduction

A smart card can be viewed as a secure data container, since it securely stores data and it is securely used during short transactions. Its safety relies first on the underlying hardware. To resist probing an internal bus, all components (memory, CPU, cryptoprocessor...) are on the same chip which is embedded with sensors covered by a resin. Such sensors (light sensors, heat sensors, voltage sensors, etc.) are used to disable the card when it is physically attacked. The software is the second security barrier. The embedded programs are usually designed neither for returning nor modifying sensitive information without guaranty that the operation is authorized.

Java Card is a kind of smart card that implements the standard Java Card 3.0 [7] in one of the two editions "Classic Edition" or "Connected Edition". Such smart card embeds a virtual machine, which interprets application bytecodes already romized with the operating system or downloaded after issuance. Due to security reasons, the ability to download code into the card is controlled by a protocol defined by Global Platform [4]. This protocol ensures that the owner of the code has the necessary credentials to perform the action. Java Cards have

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 459–468, 2010.

shown an improved robustness compared to native applications regarding many attacks. They are designed to resist to numerous attacks using both physical and logical techniques. Currently, the most powerful attacks are hardware based attacks and particularly fault attacks. A fault attack modifies part of memory content or signal on internal bus and lead to deviant behaviour exploitable by an attacker. A comprehensive consequence of such attacks can be found in [6]. Although fault attacks have been mainly used in the literature from a cryptanalytic point of view see [1, 5, 8], they can be applied to every code layers embedded in a device. For instance, while choosing the exact byte of a program the attacker can bypass countermeasures or logical tests. We called such modified application mutant.

Designing efficient countermeasures against fault attacks is important for smart card manufacturers but also for application developers. For the manufacturers, they need countermeasures with the lowest cost in term of memory and processor usage. These metrics can be obtained with an evaluation on a target [9]. For the application developers, they have to understand the ability of their applets to become mutants and potentially hostile in case of fault attack. Thus the coverage (reduction of the number of mutant generated) and the detection latency (number of instructions executed between an attack and its detection) are the most important metrics. In this paper we present a workbench to evaluate the ability of a given application to become a hostile applet with respect to the different implemented countermeasures, and the fault hypothesis.

The rest of this paper is organized as follow: first, we introduce a brief state of the art of fault injection attacks and existing countermeasures, then we discuss about the new countermeasure we have developed.Then, we present the experimentation and the results, and finally we conclude with the perspectives.

2 Fault Attacks

Faults can be induced into the chip by using physical perturbations in its execution environment. These errors can generate different versions of a program by changing some instructions, interpreting operands as instructions, branching to other (or invalid) labels and so on.

To prevent a fault attack to happen, we need to know what its effects on the smart card are. References [3, 13] has already discussed about fault model in detail.

In real life, an attacker physically injects energy in a memory cell to change its state. Thus and up to the underlying technology, the memory physically takes the value 0x00 or 0xFF. If memories are encrypted the physical value becomes a random value (more precisely a value which depends on the data, the address, and an encryption key). To be as close as possible to the reality, we have decided to choose the *precise byte error* that is the most realistic fault model. Thus, we assume that attacker can:

- make a fault injection at a precise clock cycle (she can target any operation she wants),
- only set or reset a byte to 0x00 or to 0xFF up to the underlying technology (bsr fault type), or she can change this byte to a random value out of his control (random fault type),
- target any memory cell she desires (she can target a precise variable or register).

3 Defining a Mutant Application

The mutant generation and detection is a new research field introduced simultaneously by [2, 12] using the concepts of combined attacks, and we have already discussed about mutant detection in [11]. To define a mutant application, we use an example on the following debit method that belongs to a wallet Java Card applet. In this method, the user pin must be validated prior to the debit operation.

```
private void debit(APDU apdu) {
  if ( pin.isValidated() ) {
     // make the debit operation
  } else {
     ISOException.throwIt (SW_PIN_VERIFICATION_REQUIRED);
  }
}
```

In Table 1 resides the corresponding bytecode representation. An attacker wants to bypass the pin test. She injects a fault on the cell containing the conditional test bytecode. Thus the `ifeq` instruction (byte 0x60) changes to a `nop` instruction (byte 0x00). The obtained Java code follows with its bytecode representation in Table 2.

```
private void debit(APDU apdu) {
     // make the debit operation
     ISOException.throwIt (SW_PIN_VERIFICATION_REQUIRED);
}
```

Table 1. Bytecode representation before attack

Byte	Bytecode
00 : 18	00 : aload_0
01 : 83 00 04	01 : getfield #4
04 : 8B 00 23	04 : invokevirtual #18
07 : 00 3B	07: ifeq 59
10 : ...	10 : ...
...	...
59 : 13 63 01	59 : sipush 25345
63 : 8D 00 0D	63 : invokestatic #13
66 : 7A	66 : return

Table 2. Bytecode representation after attack

Byte	Bytecode
00 : 18	00 : aload_0
01 : 83 00 04	01 : getfield #4
04 : 8B 00 23	04 : invokevirtual #18
07 : 00	07 : nop
08 : 00	08 : nop
09 : 3B	09 : pop
10 : ...	10 : ...
...	...
59 : 13 63 01	59 : sipush 25345
63 : 8D 00 0D	63 : invokestatic #13
66 : 7A	66 : return

The verification of the pin code is bypassed, the debit operation is made and an exception is thrown but too late because the attacker had already achieved his goal. This is a well example of dangerous mutant application: "*an application that has been modified by an attack that is correct for the virtual machine interpreter but that doesn't have the same behavior than the original application*". This attack has modified the control flow of the application and the goal of the countermeasure developed in this paper is to detect when such modification happen.

4 A Novel Approach to Path Check during Application Runtime

We have already proposed several solutions to check code integrity during execution in our previous publications [9, 10]. This paper is about the control flow integrity. Thus this section discusses existing countermeasures which protect the control flow integrity.

4.1 Using Java Annotation

The proposed solution uses Java annotations, when the virtual machine interpreter encounters an annotation it switches to a "secure mode". The fragment of code that follows, shows the use of an annotation on the debit method. The `@SensitiveType` annotation denotes that this method must be checked for integrity with the `PATHCHECK` mechanism.

```
@SensitiveType{
    sensitivity= SensitiveValue.INTEGRITY,
    proprietaryValue="PATHCHECK"
}
private void debit(APDU apdu) {
    if ( pin.isValidated() ) {
    // make the debit operation
    } else {
        ISOException.throwIt(SW_PIN_VERIFICATION_REQUIRED);
    }
}
```

With this approach, we provide a tool that process an annotated classfile. The annotations become a custom component containing security information. This is possible because the Java Card specification [20] allows adding custom components to a classfile: a virtual machine processes custom components if it knows how to use them or else, silently ignores them. But to process the information contained in these custom components the virtual machine must be modified.

This approach allows that to achieve a successful attack, an attacker needs to simultaneously inject two faults at the right time, one on the application code, the other on the system during its interpretation of the code which is something hard to realize and outside the scope of the chosen fault model. Now we expose the principle of the detection mechanism.

4.2 Principle of the "PATHCHECK" (PC) Method

The principle of the mechanism is divided in two parts: one part off-card and one part on-card. Our module works on the byte code, and it has at its disposal sufficient computation power because all the following transformations and computations are done on a server (off-card). It is a generalist approach that is not dependent of the type of application. But it cannot be applied to native code such as cryptographic algorithm.

Off-card. The first step is to create the control flow graph of the annotated method (in the case that it is an annotated class the operation is repeated for all the method belonging to the class), by separating its code into basic blocks and by linking them. A basic block is a set of uninterrupted instructions; It is ended by any byte code instruction that can break the control flow of the program.

Once the method is divided into basic blocks, the second step is to compute its control flow graph; the basic blocks represent the vertices of the graph and directed edges in the graph denote a jump in the code between two basic blocks (c.f. Fig. 2).

The third step is about computing for each vertex that compounds the control flow graph a list of paths from the beginning vertex. The computed path is encoded using the following convention:

- Each path begins with the tag "01". This to avoid an attack that changes the first element of a path to 0x00 or to 0xFF.
- If the instruction that ends the current block is an unconditional or conditional branch instruction, when jumping to the target of this instruction (represented by a low edge in Fig. 2), then the tag "0" is used.

– If the execution continues at the instruction that immediately follows the final instruction of the current block (represented by a top edge in Fig. 2), then the tag "1" is used.

If the final instruction of the current basic block is a switch instruction, a particular tag is used, formed by any number of bits that are necessary to encode all the targets. For example, if we have four targets, we use three bits to code each branch (like in Fig. 1). Switch instructions are not so frequent in Java Card applications. And to avoid a great increase of the application size that uses this countermeasure, they must be avoided. Thus a path from the beginning to a given basic block is $X_0 ... X_n$ (where X corresponds to a 0 or to a 1 and n is the maximum number of bit necessary to code the path). In our example, to reach the basic block 9, which contains the update of the balance amount, the paths are : `01 0 0 0 0 0 0 1` and `01 0 0 0 0 0 1`.

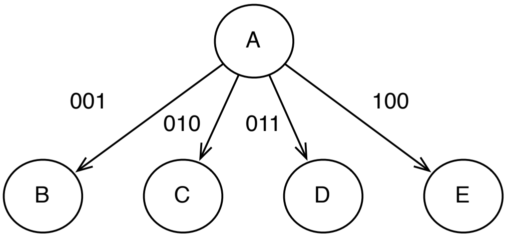


Fig. 1. Coding a switch instruction

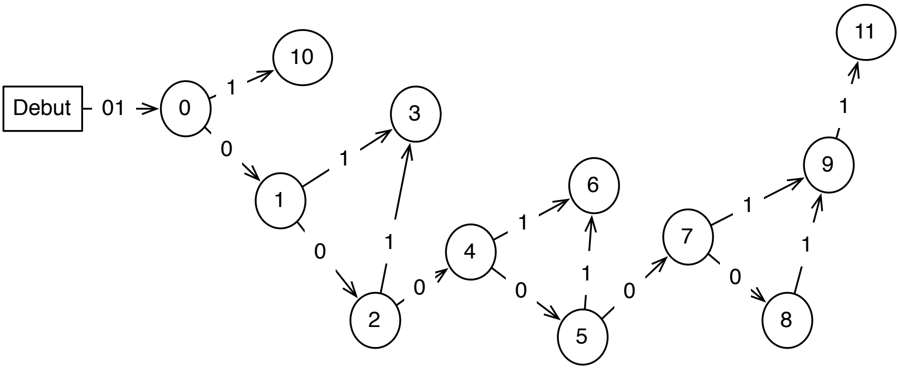


Fig. 2. Control flow graph of the debit method

On-card. When interpreting the byte code of the method to protect, the virtual machine looks for the annotation and analyzes the type of security it has to use. In our case, it is the path check security mechanism. So during the code interpretation, it computes the execution path; for example, when it encounters a branch instruction, when jumping to the target of this instruction then it saves the tag “0”, and when jumping to the instruction that follows it saves the tag “1”. Then prior to the execution of a basic block, it checks that the followed path is an authorized path i.e a path that belong to the list of path computed for this basic block. For the basic block 9, it is necessary one of the two previous paths, if not it is probably because to arrive here the interpreter has followed a wrong path; therefore, the card can lock itself.

In the case that a loop is detected (backward jump) during the code interpretation, then the interpreter checks the path for the loop, the number of reference and the number of value on the operand stack before and after the loop, to be sure that for each round the path remains the same.

5 Experimentation and Results

5.1 Resources Consumption

Table 3 shows the metrics for resources consumption obtained by activating the detection mechanism on all the method of our test applications. The increasing of the application size is variable, this is due to the number of paths presents on a method. Even if the mechanism is close to 10 % increasing of application size and 8 % of CPU overhead, the developer can choose when to activate only for sensitive methods to preserve resources. This countermeasure needs small changes on the virtual machine interpreter if we refer to the 1 % of increasing. So we can conclude that it is an affordable countermeasure.

Table 3. Ressources consumption

Countermeasures	EEPROM	ROM	CPU
Field of bits	+ 3 %	+ 1 %	+ 3 %
Basic block	+ 5 %	+1 %	+ 5 %
Path check	+ 10 %	+1 %	+ 8 %

5.2 Mutant Detection and Latency

To evaluate the path check detection mechanism, we have developed an abstract Java Card virtual machine interpreter. This abstract interpreter is designed to follow a method call graph, and for each method of a given Java Card applet, it simulates a Java Card method’s frame. A frame is a memory area allocated for the operand stack and the local variables of a given method.

The interpreter can also simulate an attack by modifying the method’s byte array. This is important because it allows to reproduce faults on demand. On top of the abstract interpreter, we have developed a mutant generator. This tool can generate all the mutants corresponding to a given application according to the chosen fault model. To realize this, for a given opcode, the mutant generator changes its value from 0x00 to 0xFF, and for each of these values an abstract interpretation is made. If the abstract interpretation does not detect a modification then a mutant is created enabling us to regenerate the corresponding Java source file and to color the path that lead to this mutant.

The mutant generator has different mode of execution:

- *The basic mode*: the interpreter executes the instruction pushing and popping element on the operands stack and using local variables without check. In this configuration instructions can use elements of other methods frame like using their operands stack or using their locals. When running this mode, it has no countermeasures activated.
- *The simple mode*: the interpreter checks that no overflow or no underflow occurs, that the used locals are inside the current table of locals, and that when a jump occurs it's done inside the method. They consist in some verifications done by the Java verifier.
- *The advanced mode*: is the simple mode with the ability to activate or to deactivate a given countermeasures like the developed ones:path checking mechanism (PC), field of bits mechanism (FB) see [9], or PS mechanism. PS is a detection mechanism that is not described in this paper and for which a patent is pending.

The Table 4 shows the reduction of generated mutants in each mode of the mutant generator for an application. The second line shows the number of mutant generated in each mode of the mutant generator. The third line of those tables shows the latency.

The latency is the number of instruction executed between the attack and the detection. In the basic mode no latency is recorded because no detection is made. This value is also really important because if a latency if too high maybe instructions that modify persistent memory like: `putfield`, `putstatic` or an invoke instruction (`invokestatic`, `invokevirtual`, `invokespecial`, `invokeinterface`) can be executed. If a persistent object is modified then it is manipulated during all the future session between the smart card and a server. So this value has to be as small as possible to lower the chances to have instructions that can modify persistent memory.

Table 4. Wallet (simple class) - 470 Instructions

	Basic mode	Simple mode	PC	FB	PS
Number of mutants	440	54	23	10	30
Latency	-	2,91	3,33	2,43	2,92

Path check fails to detect mutant when the fault that generate the mutant don't influence the control flow of the code. Otherwise, when a fault occurs that alter the control flow of the application then this countermeasure detects it. With this countermeasure it becomes impossible to bypass systems calls like cryptographic keys verification. And if it remains some mutant, applicative countermeasures can be applied on demand to detect them.

6 Conclusion and Future Works

We had presented in this paper, a new approach that is affordable for the card and that is fully compliant with the Java Card 2.x and 3.x specification. Moreover it does not consume too much computation power and the produced binary files are under a reasonable limit in term of size. It does not disturb the applet conception workflow, because we just add a module that will makes lightweight modification of the byte code. It saves time to the developer who wants to produce secured applications thanks to the use of the sensitive annotation. Finally, it needs a tiny modification of the java virtual machine. It also has a good mutant applications detection capacity.

We have implemented all these countermeasures inside a smart card in order to have metrics concerning memory footprint and processor overhead, which are all affordable for smart card. In this paper we presented the second part of this characterization to evaluate the efficiency of countermeasures in smart card operating system. We provide a framework to detect mutant applications according to a fault model and a memory model. This framework is able to provide to a security evaluator officer all the source code of the potential mutant of the application. She can decide if there is a threat with some mutants and then to implement a specific countermeasure.

Within this tool, either the developer or security evaluator officer is able to take adequate decision concerning the security of its smart card application. For the developer company, reducing the size of the embedded code minimizes the cost of the application. For the security evaluator it provides a semi automatic tool to perform vulnerability analysis.

References

[1] Aumuller, C., Bier, P., Fischer, W., Hofreiter, P., Seifert, J.: Fault attacks on RSA with CRT: Concrete results and practical countermeasures. LNCS, pp. 260–275 (2003)

[2] Barbu, G., Thiebeauld, H., Guerin, V.: Attacks on Java Card 3.0 Combining Fault and Logical Attacks. In: Gollmann, D., Lanet, J.-L., Iguchi-Cartigny, J. (eds.) CARDIS 2010. LNCS, vol. 6035, pp. 148–163. Springer, Heidelberg (2010)

[3] Blomer, J., Otto, M., Seifert, J.P.: A new CRT-RSA algorithm secure against Bellcore attacks. In: Proceedings of the 10th ACM conference on Computer and communications security, pp. 311–320. ACM, New York (2003)

[4] Global platform group. Global platform official site (July 2010) http://www.globalplatform.org

[5] Hemme, L.: A differential fault attack against early rounds of (triple-) DES. In: Joye, M., Quisquater, J.-J. (eds.) CHES 2004. LNCS, vol. 3156, pp. 254–267. Springer, Heidelberg (2004)

[6] Iguchi-Cartigny, J., Lanet, J.L.: Developing a Trojan applets in a smart card. Journal in Computer Virology, 1–9 (2009)

[7] Sun Mycrosystems, Java CardTM 3.0.1 Specification. Sun Microsystems (2009)

[8] Piret, G., Quisquater, J.-J.: A differential fault attack technique against SPN structures, with application to the AES and Khazad. In: Walter, C.D., Koç, Ç.K., Paar, C. (eds.) CHES 2003. LNCS, vol. 2779, pp. 77–88. Springer, Heidelberg (2003)

[9] Sere, A.A., Iguchi-Cartigny, J., Lane, J.L.: Automatic detection of fault attack and countermeasures. In: Proceedings of the 4th Workshop on Embedded Systems Security, pp. 1–7. ACM, New York (2009)

[10] Sere, A.A., Iguchi-Cartigny, J., Lanet, J.L.: Checking the Path to Identify Control Flow Modification. PAca Security Trends In embedded Systems (2010)

[11] Sere, A.A., Iguchi-Cartigny, J., Lanet, J.L.: Mutant applications in smart card. In: Proceedings of CIS 2010(2010)

[12] Vetillard, E., Ferrari, A.: Combined Attacks and Countermeasures. In: Gollmann, D., Lanet, J.-L., Iguchi-Cartigny, J. (eds.) CARDIS 2010. LNCS, vol. 6035, pp. 133–147. Springer, Heidelberg (2010)

[13] Wagner, D.: Cryptanalysis of a provably secure crt-rsa algorithm. In: Proceedings of the 11th ACM conference on Computer and communications security, pp. 92–97. ACM, New York (2004)

Enhanced Sinkhole System by Improving Post-processing Mechanism

Haeng-Gon Lee, Sang-Soo Choi, Youn-Su Lee, and Hark-Soo Park

Science and Technology Security Center (S&T-SEC),
Korea Institute of Science and Technology Information (KISTI),
Daejon, 305-806, Korea
{hglee,choiss,zizeaz,hspark}@kisti.re.kr

Abstract. Cybercrime is threatening our lives more seriously. In particular, the botnet technology is leading most of cybercrime such as distribute denial of service attack, spamming, critical information disclosure. To cope with this problem, various security techniques have been proposed. Especially, DNS-Sinkhole is known as the most effective approach to detect botnet activities. It has various advantages such as low cost, easy establishment and high effect. However, botnet response is more difficult because botnet technology is constantly evolving. In particular, legacy sinkhole system has revealed a variety of limitations such as low accuracy and limited information. Therefore, additional research is required to overcome these limitations. In this paper, we propose an enhanced sinkhole system that utilizes DNS-Sinkhole. Especially, we focus on the improving of post-processing mechanism based on packet analysis.

Keywords: Botnet defense, DNS-Sinkhole based bot response, packet analysis.

1 Introduction

This century has become "Information Age" by the rapid progress of IT technology and widespread Internet usage. On information oriented society, people can enjoy convenient life such as online shopping, banking and studying with less time-wasting and small effort.

However, it doesn't imply a rosy future. It comes with not only good effects but also with bad effects. Various cyber attacks threaten our privacy and causes financial losses. Especially, bot and botnet technology have emerged as a hot issue in the field of information security [1]. These are useful for hackers who want to perform a variety of cyber attacks such as spam-mail sending, distribute denial of service (DDoS) attack. Moreover, complete countermeasure which can fundamentally prevent bot and botnet activities does not exist yet [2].

The DDoS attack which started at July 7th 2009 is one of the most serious cases. It had been performed with 30~180 thousand bots for three days. Security officials had to struggle to clean up each bots (especially, zombies), but it was inadequate to protect 35 important organizations' websites [3]. In addition, by the remaining bots, the attack recurred at the same day of 2010. Lessons from this case are clear: we need an effective solution to protect our information assets from botnet attack.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 469–480, 2010.

One of the possible solutions against botnet attack is DNS-Sinkhole technique. It blocks zombies trying to connect to command and control (C&C) server by adapting simple configuration setting on DNS server. Especially, it is known to be most effective C&C based botnet detection technique. However, the needs for improving DNS-Sinkhole technique have increased, because botnet technology is developing rapidly. To cope with this problem, we propose an enhanced sinkhole system. In particular, we focus on improving post-process mechanisms of DNS-Sinkhole.

This paper is organized as follows. Introduction and analysis of advantages as well as limitations of DNS-Sinkhole technique are described in chapter 2. The proposed post-processing mechanism for DNS-Sinkhole and enhanced sinkhole system are covered in chapter 3. Analysis of operation results by using the enhanced sinkhole system is presented in chapter 4. Finally, chapter 5 concludes the paper.

2 Related Work

DNS-Sinkhole is one of the botnet countermeasures, and it forces command and control messages between bots and C&C server to detour toward itself. Consequently, numerous bots that require IP address from the DNS server linked with DNS-Sinkhole are unable to perform actual attacks. So far, it is known as the best practice of botnet countermeasures.

2.1 DNS-Sinkhole Technique

DNS-Sinkhole is one of the botnet countermeasures, and it forces command and control messages between bots and C&C server to detour toward itself. Consequently, numerous bots that require IP address from the DNS server linked with DNS-Sinkhole are unable to perform actual attacks. So far, it is known as the best practice of botnet countermeasures [4, 5].

Fig.1 presents basic principle of DNS-Sinkhole. In particular, it has a two-tier structure. One is the pre-processing technique which focuses on connection redirection by adapting blacklists. The other is the post-processing technique which focuses on connection detection by monitoring IRC/HTTP daemons.

Most of the botnet use C&C server for sending commands to bots. To avoid IP based information security systems such as firewall and intrusion detection system, bots which are infected systems by malware to receive orders from the bot-master store C&C server's information: not IP address but domain name. After getting IP address by sending DNS query (the requirement of domain name) to DNS server, bots can access to C&C server. If bots don't receive C&C server's IP address, they can't communicate with bot-master and take any attack action. Following is the process of DNS query:

① Bots require C&C server's IP address.
② DNS server send the IP address.
③ Bots access to C&C server.
④ C&C server send commands to bots.

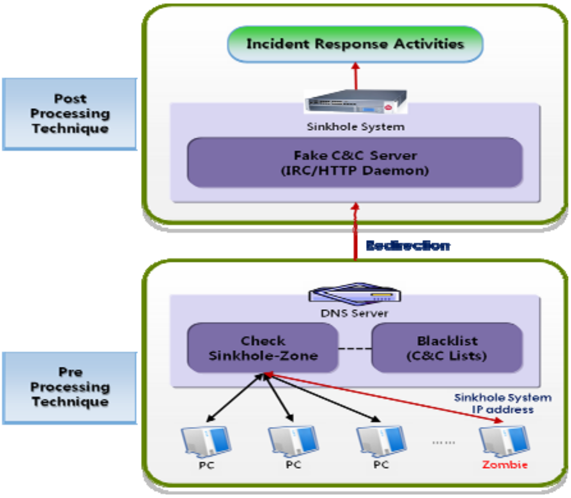


Fig. 1. Basic principle of DNS-Sinkhole

DNS-Sinkhole interrupts command and control schemes between bots and C&C server by absorbing DNS queries of bots: The Sinkhole-Action. Following is the process of DNS query occurred from the DNS server applied DNS-Sinkhole:

① Bots require C&C server's IP address.
② DNS server send sinkhole server's IP address.
③ Bots access to DNS-Sinkhole server.
④ No attack action is performed by bots.

For the Sinkhole-Action, DNS server refers to "Blacklist" in its first BIND configuration file. The Blacklist contains malicious domain names configured to find Sinkhole server's zone file. Fig.2 will be helpful to understand how to operate the Sinkhole-Action in DNS server.

Among the botnet solutions, the DNS-Sinkhole is estimated like the best practice. It shuts commands out of numerous bots at one time. In addition, the cost of the system establishment is very low. The only required payment is buying a few servers. If you have spare systems, additional costs are absolutely zero.

2.2 Advantages and Limitations

DNS-Sinkhole is, without doubt, very effective to protect own systems from being used for botnet attacks. In particular, it is the most cost-effective botnet solution. However, the security environment is changed rapidly and DNS-Sinkhole has begun to reveal the following limitations.

Limited Information. In nowadays, the techniques of cyber attack are getting too complicated. Most of C&C server used IRC protocol with the central structure in the past. However, botnets use various protocols such as IRC, HTTP, P2P, etc as shown

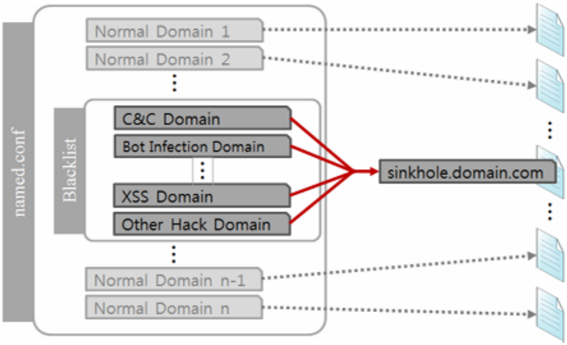
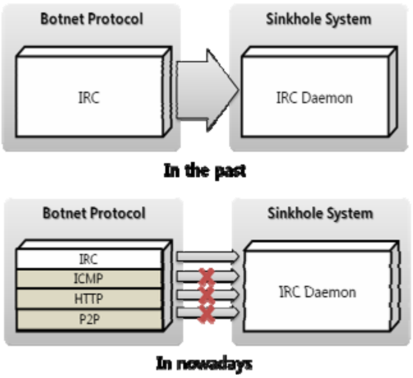


Fig. 2. Sinkhole-Action in DNS server

in Fig.3. Existing DNS-Sinkhole server (proposed by KISA in 2004, Korea) has an IRC daemon to analyze incoming botnet (especially, zombies which are infected by malicious code) information. Though the analyzing isn't connected with the protecting action directly, it is important to build a prevention strategy. This system must improve analysis capacity to adjust changed environment.

Lack of Connection with Incident Response. The Sinkhole-Action doesn't contain any action for removing the malicious code. Although the botnet attack is blocked by the system in advance, the network still contains the risk of recurrence until clean up malicious codes from zombies because there are no actions for removing as shown in Fig.4. The recurrence of July 2009 cyber attacks at 7th July 2010 is a good example. The main problem is still living zombies. Considerations for connection with actual incident response activities are strongly required. In particular, actual removal of malicious code on zombies can be achieved by incident response activities.



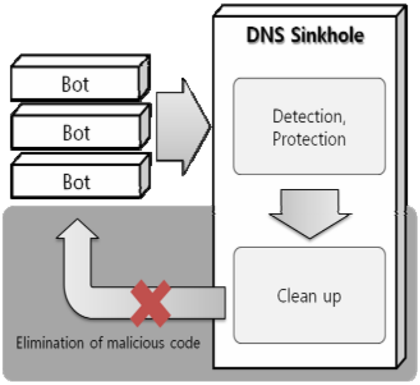


Fig. 3. The evolution of botnet protocols **Fig. 4.** Absence of cleaning up activities

Low Accuracy. DNS server classifies incoming queries, whether it is a malicious domain or not, by comparing with blacklists. In particular, it is made by provider and is updated periodically to DNS server. If some domain is judged as a malicious domain, the system absorbs the whole requiring from PCs. However, there are two problems as shown in Fig.5: reliability of the blacklist and sub-domain. When the blacklist has wrong items in its list, or some system (listed in blacklists) cleans malicious code without the recognition of provider, the list is not perfectly correct. However, on DNS-Sinkhole system, blacklist is trusted without doubt, because there is no process for verification. In addition, it doesn't check sub-domain. The judgment bases on primary domain name only. On hosting service environment, incorrect detection is possible frequently. The improvement of accuracy is important for user convenience and the system reliability.

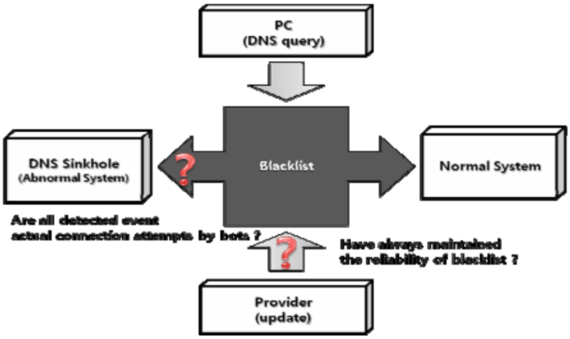


Fig. 5. Accuracy problem of the blacklists

To cope with those problems, we propose the enhanced sinkhole system by improving post-processing mechanism of DNS-Sinkhole. Proposed system should be satisfied with three requirements that come from the limitations as followings:

① Improving of information gathering process
② Considering of connection with incident response activities
③ Increasing of detection accuracy rate.

3 Enhanced Sinkhole System

Fig.6 is the structure and process of enhanced sinkhole system (ESS). In particular, it is based on packet analysis for the purpose of improving post-processing mechanism of DNS-Sinkhole.

When PCs send DNS query, DNS server figures out whether it includes malicious domain or not by blacklist, and sends it to the ESS. ESS has listening daemons to receive packets from HTTP and other protocols.

Especially, ESS consists of three modules: analyzer, rule checker, dedicated database. Analyzer module performs information gathering process. To increase detection accuracy rate, rule checker module classifies incoming data whether it is truly related with botnet or not. Dedicated database stores a series of data which are reformed by analyzer module and classified by rule checker module. In particular, incident response team can use this information to mitigate or clean up effectively focused on botnet.

3.1 Analyzer Module

Analyzer module is a kind of packet analysis module. The purpose of the module is to extract significant information for botnet analysis. Botnet analysis is necessary to figure out specific victim machine (zombie PC), time that malicious action is occurred, and types of botnet. From incoming packets, analyzer module extracts the information related above.

Analyzer module checks the packets' capacity of analysis, and then decides type of protocol, after then makes suitable information for each protocol as shown in Fig.7. Input data should also be able to be interpreted by analyzer module, because ESS is designed for practical use.

In the first process, analyzer module checks packets whether they contain the readable data (character, number, and special character) more than 50%. If the packet is accepted, it goes to next process. Usually, encrypted or damaged packets are excluded. Fig.8 presents a sample of unaccepted packet.

In the second process, protocol type of packets is inspected. Each packet has different structure depending on the type of protocol. Therefore, required data for detecting botnet is different too.

PacketType() is one of functions in analyzer module that performs first and second process as shown in Fig.9.

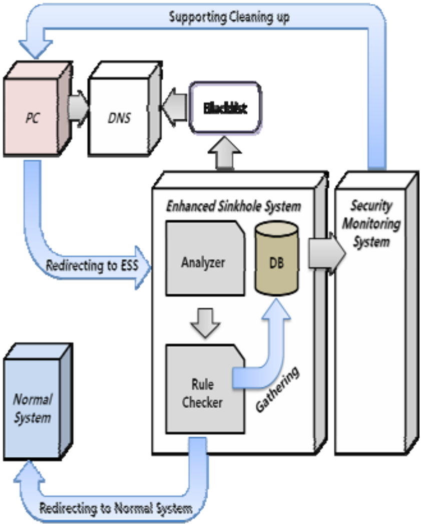


Fig. 6. Design of proposed ESS

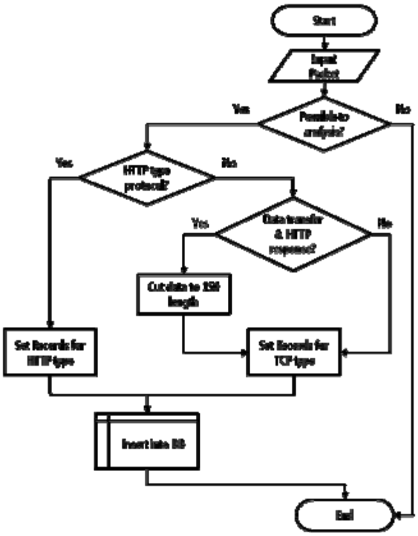
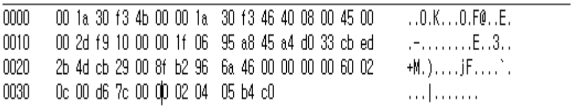


Fig. 7. Flow chart of analyzer module



```
0000   00 1a 30 f3 4b 00 00 1a  30 f3 46 40 08 00 45 00    ..0.K...0.F@..E.
0010   00 2d f9 10 00 00 1f 06  95 a8 45 a4 d0 33 cb ed    .-........E..3..
0020   2b 4d cb 29 00 8f b2 96  6a 46 00 00 00 00 60 02    +M.)....jF....`.
0030   0c 00 d6 7c 00 00 02 04  05 b4 c0                   ...|.......
```

Fig. 8. Unaccepted packet by analyzer module

```
public int packetType(byte[] data)
{
  if /* packet is available to analysis */
  {
    // Process for valid data
    String data1 = Encoding.ASCII.GetString(data);
    result = 1; // TCP type

    // Process to check Protocol Type
    String tmp = Encoding.ASCII.GetString(data).
                    Substring(0,4);
    if (tmp.Equals("GET ") || tmp.Equals("POST"))
      result = 2; // HTTP type
    else result = 3; // un-avaliable to a nalysis (drop)
  return result;
  }
}
```

Fig. 9. PacketType( ) function

Table 1 presents an example of HTTP based bot and its request. In general, bots send requests with a consistent pattern to connect to C&C server. If patterns are analyzed, the decision making is possible whether it is related with botnet action.

Fig.10 shows PacketAnalysis() function which collects necessary data from packets. In particular, it collects detail data to judge detection accuracy of HTTP botnet: Host, User-Agent, Get value, and etc. Therefore, it checks protocol type in advance. Finally, collected data is written on database.

Table 1. Bot samples with request

Trojan-Downloader.Win32.Agent.dico [Kaspersky Lab]	
Request	http://aahydrogen.com/maczjwtq/iolylzjjg.php?adv=adv441 http://aahydrogen.com/maczjwtq/birqakky.php?adv=adv441 http://aahydrogen.com/maczjwtq/yekhhiijfg.php?adv=adv441 http://bastocks.com/maczjwtq/iolylzjjg.php?adv=adv441 http://bastocks.com/maczjwtq/birqakky.php?adv=adv441 http://bastocks.com/maczjwtq/yekhhiijfg.php?adv=adv441
Trojan-Downloader.Win32.Zlob.bgs [Kaspersky Lab]	
Request	http://www.bqgate.com/gatech.php?pn=srch0p1total7s2 http://www.bqgate.com/index.jpg http://www.bqgate.com/gatech.php?id=dw01 http://www.bqgate.com/gatech.php?pn=srch0p2total7s2 http://www.bqgate.com/gatech.php?pn=srch0p3total7s2 http://www.bqgate.com/gatech.php?pn=srch0p4total7s2 http://www.bqgate.com/gatech.php?pn=srch0p10total7s2 http://www.bqgate.com/gatech.php?pn=srch0p11total7s2

3.2 Rule Checker Module

Rule checker module analyzes correlations of botnets with incoming packets for improving detection accuracy of ESS. One of critical issues of legacy system is blocking

```
public PacketData packetAnalysis(TCPPacket tcp)
{
  int type = packetType(tcp.Data);

  if (type == 2) //HTTP packet
  {
    for (int i = 0; i < payload.Length; i++)
    {
      if (/** enough to split **/)
      payload[i] = payload[i].Substring(0,(payload[i].
                     Length - 1));
    }
    processPacket.setGetValue(payload[0]);
    foreach (String pkt in payload)
    {
      processPacket.setHost(pkt);
      processPacket.setUserAgent(pkt);
    }
  } else if(type == 1) //other TCP packet
  {
    String tmpdata =Encoding.Default.GetString(tcp.TCPData);
    if (tmpdata.Length > 150 && tmpdata.Substring(0,7).
        Equals("HTTP/1."))
      tmpdata = tmpdata.Substring(0, 150);
      processPacket.setValue(tmpdata.Replace("'", "\'"));
  }
  return processPacket;
}
```

Fig. 10. PacketAnalysis( ) function

whole access to the specific domain. If some portal site is exploited as a detour or sub-domain is used for malicious action, legacy system blocks whole access to them. It disturbs user convenience and decreases reliability.

To cope with this problem, rule checker module compares the information received from analyzer module with signature group, then judges whether the detection is correct or not. If it is classified as wrong detection, ESS redirect packet to normal domain. In case of accurate detection, information is stored to dedicated database. Signature has types divided by protocol type and it is defined by analyzing various botnets.

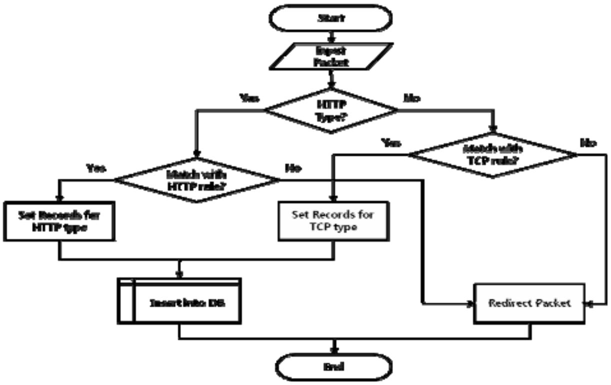


Fig. 11. Flow chart of rule checker module

Fig.11 presents flow chart for Rule Checker. Incoming packet's protocol type is classified at first process, and then it is compared with selected signature group. If matching data is exist, it is defined as a correct detection and information is stored to DB. If it is defined as a wrong detection, the packet is redirected to normal destination address.

As mentioned above, signatures used in rule checker module are made by analyzing botnets. Table 2 is a sample signature by analyzing bots introduced on Table 1. In particular, three types of information (domain name, get value and user agent) should be detected for detailed analysis of HTTP bots.

Table 2. Sample signature for checking rule

Trojan-Downloader.Win32.Agent.dico [Kaspersky Lab]	
Host	http://aahydrogen.com/ http://bastocks.com/
GetValue	*.php?adv=adv441
User Agent	[no matter]
Trojan-Downloader.Win32.Zlob.bgs [Kaspersky Lab]	
Request	http://www.bqgate.com/
GetValue	gatech.php?pn=srch[*]p1total7s2
User Agent	[no matter]

3.3 Dedicated Database

Information analyzed and collected by ESS is very useful for incident response activities. Especially, we used this information to security monitoring system (SMS) for the purpose of cleaning up malicious codes and updating signatures.

ESS has three database tables as shown in Fig.12. Especially, these tables are connected with security monitoring service as shown in Fig.13.

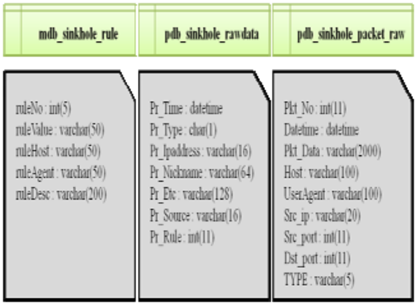


Fig. 12. DB tables for ESS

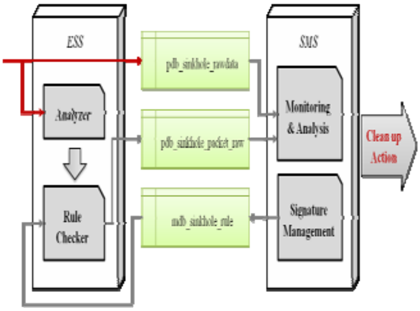


Fig. 13. Data sharing between ESS and SMS

4 Analysis of Operation Results

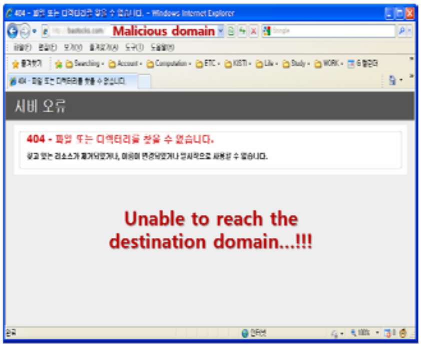
We have been operating legacy sinkhole system in January 2010, and we applied ESS in June 2010 as shown in Table 3. Therefore, we compare legacy system and proposed ESS by analyzing whether the limitations are solved.

Table 3. Operation environment of ESS

Items	Description	Note
Monitoring Target	- 25 DNS server	Private DNS
Server Spec	**3 Servers** - Sinkhole System (2) - Database Server (1)	NLB
Period	- About 3 months	

If one system tries to connect to malicious domain, it receives a response as shown in Fig.14. In this case, legacy sinkhole system receives detection event logs which are generated by IRC daemon as shown in Fig.15. Especially, proposed ESS performs detailed analysis based on packet data to improve legacy system.

First, ESS gathers more information. Table 4 presents the comparison of legacy system and proposed ESS. The scope of protocol is enlarged from only IRC to IRC, HTTP, and other TCP protocols. In addition, the scope of supporting information is enlarged also. Though legacy system supports only three items, ESS supports six items.



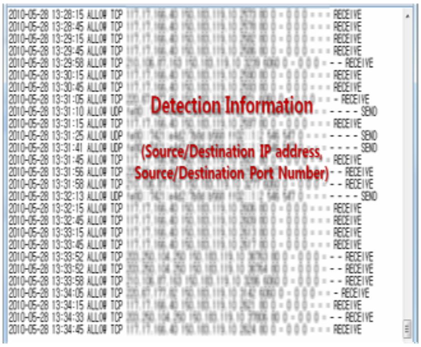
404 - 파일 또는 디렉터리를 찾을 수 없습니다. - Windows Internet Explorer

서버 오류

404 - 파일 또는 디렉터리를 찾을 수 없습니다.

찾고 있는 리소스가 제거되었거나, 이름이 변경되었거나 일시적으로 사용할 수 없습니다.

Fig. 14. Result for request to malicious domain



```
2010-05-28 13:28:15 ALLOW TCP [redacted] RECEIVE
2010-05-28 13:28:45 ALLOW TCP [redacted] RECEIVE
2010-05-28 13:29:15 ALLOW TCP [redacted] RECEIVE
2010-05-28 13:29:45 ALLOW TCP [redacted] RECEIVE
2010-05-28 13:29:58 ALLOW TCP [redacted] - - RECEIVE
2010-05-28 13:30:15 ALLOW TCP [redacted] RECEIVE
2010-05-28 13:30:45 ALLOW TCP [redacted] RECEIVE
2010-05-28 13:31:05 ALLOW TCP [redacted] - RECEIVE
2010-05-28 13:31:10 ALLOW UDP [redacted] - - - - - SEND
2010-05-28 13:31:15 ALLOW TCP [redacted] RECEIVE
2010-05-28 13:31:25 ALLOW UDP [redacted] - - - - - SEND
2010-05-28 13:31:41 ALLOW UDP [redacted] - - - - - SEND
2010-05-28 13:31:45 ALLOW TCP [redacted] RECEIVE
2010-05-28 13:31:56 ALLOW TCP [redacted] - - RECEIVE
2010-05-28 13:31:58 ALLOW TCP [redacted] - - RECEIVE
2010-05-28 13:32:13 ALLOW UDP [redacted] - - - - - SEND
2010-05-28 13:32:15 ALLOW TCP [redacted] RECEIVE
2010-05-28 13:32:45 ALLOW TCP [redacted] RECEIVE
2010-05-28 13:33:15 ALLOW TCP [redacted] RECEIVE
2010-05-28 13:33:45 ALLOW TCP [redacted] RECEIVE
2010-05-28 13:33:52 ALLOW TCP [redacted] - - RECEIVE
2010-05-28 13:33:52 ALLOW TCP [redacted] - - RECEIVE
2010-05-28 13:33:58 ALLOW TCP [redacted] - - RECEIVE
2010-05-28 13:34:05 ALLOW TCP [redacted] - RECEIVE
2010-05-28 13:34:15 ALLOW TCP [redacted] RECEIVE
2010-05-28 13:34:33 ALLOW TCP [redacted] - - RECEIVE
2010-05-28 13:34:45 ALLOW TCP [redacted] RECEIVE
```

Detection Information
(Source/Destination IP address,
Source/Destination Port Number)

Fig. 15. Event log generated by legacy system

For example, it can give more detail information on report. Table 5 shows information lists that can come from two systems. System manager who received the detail report from security monitoring center can take botnet cleaning up action.

Second, botnet cleaning up is possible by offering botnet detection information to the linked monitoring system as shown in Fig.16. Real-time detection and incident response activities can be performed more effectively. In general, botnet cleaning up activities use anti-virus solutions. ESS can offer detail information about anti-virus solution lists which can eliminate malicious code on zombies. Followings are incident response process based on SMS:

① Detect a zombie based on ESS
② Analyze possible anti-virus solution based on information of ESS
③ Send a report and possible anti-virus solution lists to actual system manager
④ Clean up the system by using anti-virus solution

Table 4. Comparison of scope and information

Items	Legacy system	Proposed ESS
Protocol	- IRC	- IRC - HTTP - other TCP
Information	- Source IP - Source Port - Protocol Type	- Source IP - Source Port - Protocol Type - User Agent - Host - Request Info

Table 5. Comparison of supporting information

Information	Legacy system	Proposed ESS
Zombie	Possible	Possible
Details of attack	Impossible	Possible
Malicious Domain	Impossible	Possible
Required files	Impossible	Possible

Third, the detection accuracy is noticeably increased. Because legacy system doesn't have any process to verify whether the detection is correct or not, possibility of incorrect detection exist.

For example, legacy system detected 8,315 connections that are related with one website in three months. This website was included on the blacklists because it had one malicious webpage which were used for C&C. In particular, it included malicious file which can offer zombies to attacking command. However, only 55 connections are related with actual malicious webpage that are analyzed by ESS based on packet analysis as shown in Fig.17.

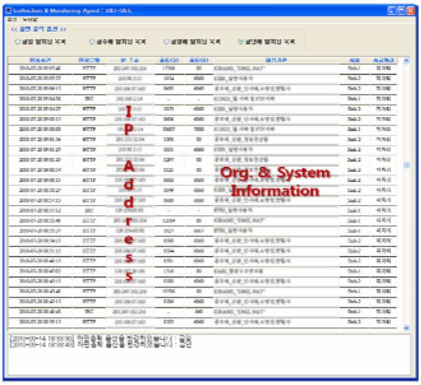


Fig. 16. Screen shot of real-time SMS

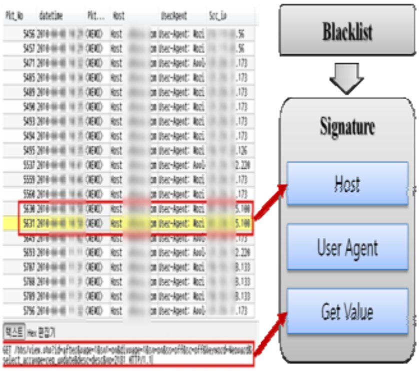


Fig. 17. Classification of connection by ESS

5 Conclusion

In this paper, we propose enhanced sinkhole system by improving post-processing mechanism of DNS-Sinkhole. Especially, proposed system overcomes the limitations which are derived from operating on legacy sinkhole system. In addition, it succeeds various advantages of general DNS-Sinkhole technique such as easy establishment, low cost. In addition, it has the following features:

① Extended detection scope (IRC, HTTP and other TCP protocols)
② Rich information based on packet analysis
③ Verification process for the blacklists
④ Redirecting connection to normal system

References

1. Ianelli, N., Hackworth, A.: Botnet as a vehicle for online crime. CERT. Request for Comments (RFC) 1700 (December 2005)
2. Bailey, M., Cooke, E., Jahanian, F., Xu, Y., Karir, M.: A Survey of Botnet Technology and Defenses. In: Proceedings of Cybersecurity Applications & Technology Conference For Homeland Security (CATCH), pp. 299–304 (2009)
3. Korea Internet & Security Agency, A Strategy and Policy Planning for DDoS Response, KISA homepage (2010)
4. Kim, Y.-B., Youm, H.-Y.: A New Bot Disinfection Method Based on DNS Sinkhole. Journal of KIISC 18(6A), 107–114 (2008)
5. Kim, Y.-B., Lee, D.-R., Choi, J.-S., Youm, H.-Y.: Preventing Botnet Damage Technique and It's Effect using Bot DNS Sinkhole. Journal of KISS(C): Computing Practices 15(1), 47–55 (2009)

A Practical Methodology and Framework for Comprehensive Incident Handling Focused on Bot Response

Sang-Soo Choi, Myung-Jin Chun, Youn-Su Lee, and Hyeak-Ro Lee

Science and Technology Security Center (S&T-SEC),
Korea Institute of Science and Technology Information (KISTI),
Daejon, 305-806, Korea
{choiss,bluebary,zizeaz,leehr}@kisti.re.kr

Abstract. Cybercrime has emerged as a major social problem. Especially, widely distributed bots are being used as a main tool for conducting cybercrime. Therefore, the needs for enhanced incident handling have increased to detect, analyze and respond to bots and botnets. DNS-Sinkhole is known as the most effective way to respond to bot activity. Incident handling is a very complex set of security activities including technical and managerial part. Especially, the concept of incident handling is higher than DNS-Sinkhole technique which only focuses on the technical part. Therefore, additional studies to integrate incident handling with DNS-Sinkhole technique are required. We propose systematic approach for comprehensive incident handling focused on the bot response. In particular, we propose comprehensive incident handling methodology based on DNS-Sinkhole technique, and practical incident handling framework for central incident handling team.

Keywords: Comprehensive Incident Handling, Practical Incident Handling Framework focused on Bot Response, DNS Sinkhole Technique.

1 Introduction

The biggest problem in the information society is an increase of cybercrime. Also, the purpose of hacking attacks is changed from hacker's self conspicuous to financial exploitation and cyber terrorism. The dramatic increase in cybercrime has emerged as a major and serious social problem. In particular, the essential hacking tool is the bots and botnets.

The bot is a kind of worm/virus that is installed secretly on personal user's PC and remotely controlled by hacker. PCs infected with bot are called zombies (or drones), are distributed widely throughout the world. Typically, zombies are used to perform distribute denial of service (DDoS) attack, spamming, privacy and critical data exposure.

In order to respond quickly and effectively from changes in cyber threat environment, security engineers and researchers have been conducting a variety of research regarding structure, detection, analysis and response focused bot and botnet [1, 2, 3, 4]. Especially, DNS-Sinkhole is known to be most effective command and control

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 481–492, 2010.

(C&C) based bot detection technique [5, 6, 7]. However, researches about incident handling methodology and framework enabling bot detection, analysis and response by using DNS-Sinkhole technique are insufficient.

Comprehensive incident handling is a very complex set of information security activities. In particular, detection, analysis and response phases are required to perform actual incident response activities [8, 9]. However, DNS-Sinkhole technique is only focuses on detection phase.

To cope with those problems, we propose a practical methodology and framework enabling comprehensive incident handling based on DNS-Sinkhole technique.

This paper is organized as follows. Basic concept of DNS-Sinkhole technique and roles for incident handling are described in chapter 2. The proposed methodology based on DNS-Sinkhole technique is covered in chapter 3. Actual incident handling framework focused on bot response is presented in chapter 4. Finally, chapter 5 concludes the paper.

2 Related Work

DNS-Sinkhole is a simple technique to construct an effective environment for detection and blocking of bot activity. However, incident handling is more complex and sophisticated set of information security activities including detection, analysis and response to cyber attacks. Thus, convergence of the technical method and actual incident handling activities are needed.

2.1 Basic Concept of DNS-Sinkhole Technique

The concept of DNS-Sinkhole was suggested by Korea Internet & Security Agency (KISA), has been utilized to respond to bots since 2004 [5, 6, 7]. DNS-Sinkhole technique blocks zombies trying to connect to C&C server. It allows malicious actions by zombies (such as distribute denial of service attack, privacy exposure) can be prevented in advance. Especially, DNS-Sinkhole is known to be most effective C&C based bot detection technique.

Fig.1 presents an overview of DNS-Sinkhole technique. Ultimately, hacker (especially, bot-master) will lose control for zombies. In the case of zombie's first attempt trying to connect to the C&C server, hackers can't add bot infected PC on the zombie-army list. These effects can be obtained by changing configuration settings of the DNS server. If you understand the domain name service, this is very simple task. Fig.2 shows a sample result of DNS zone setting.

Additionally, if you can prepare server system which performs the role of fake C&C, making zombie list and secondary security activities are also possible. Especially, installation of internet relay chat (IRC) daemon has been recommended.

As mentioned above, DNS-Sinkhole is simple but powerful technique to detect and prevent malicious activity by zombies. In addition, it can be applied very effectively by central incident response team for incident handling, because it can analyze bot and botnet trends cost-effectively.

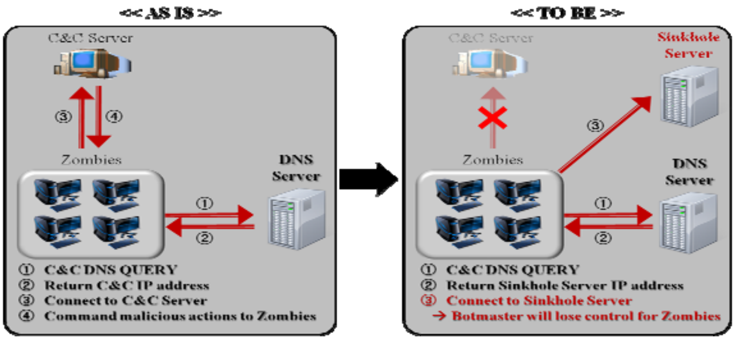


Fig. 1. Overview of DNS-Sinkhole technique

```
zone 'blacklist.net' IN { type master; file '/var/
         named/forward/blacklist/X.NET'; };
zone 'blacklist.com' IN { type master; file '/var/
         named/forward/blacklist/X.COM'; };
......
```

Fig. 2. Example of DNS setting (named.conf on Linux OS)

2.2 Basic Roles for Incident Handling

In general, an incident is a violation of computer security policies, acceptable use policies, or standard computer security practices. Incident Handling is the process of detecting and analyzing incidents and limiting the incident's effect. Essential roles in incident handling are preparation, detection / analysis, containment / eradication / recovery, and post-incident activity [8].

Security monitoring has similar concept to incident handling. It was started on network based event log monitoring. Network security monitoring is the collection, analysis and escalation of indications and warnings to detect and response to intrusions. Essential roles in network security monitoring are collection, analysis and escalation [9].

Consequently, Incident handling is a series of information security activities. However, roles of incident handling and security monitoring are too broad and abstract. In particular, individual organization and small scale of incident response team are impossible to perform all roles. Therefore, we redefine major roles for incident handling in terms of feasibility as follows: detection, analysis, and response. These are basic roles for incident handling, but additional roles will be expanded by large scale of incident response team.

Detection. In general, detection is the extraction of particular information from a larger stream of information without specific cooperation from or synchronization with the sender. In the area of information security, intrusion detection is a more generalized term. It is the act of detecting actions that attempt to compromise the confidentiality, integrity or availability of a resource.

In order to meet the detection role, following activities are required. First, security target (ST) for security monitoring must be determined. Networks, information systems and personal computers can be considered as the ST. Second, ST should be connected to the incident handling system (IHS) built in incident response team (IRT) such as security monitoring center, computer emergency response team. Especially, information security systems such as IDS, TMS, ESM can be considered as the IHS. Third, real-time monitoring should be carried out by using cyber threat information collected by IHS from ST.

The detection is the basic but very important role for successful incident handling, can be implemented in various ways. Thus, the most appropriate model should be selected through careful analysis of ST's computing environment.

Analysis. In general, analysis is the process of breaking a complex topic or substance into smaller parts to gain a better understanding of it. In the area of information security, log analysis is a more generalized term. It is an art and science seeking to make sense out of computer-generated records (also called log or audit trail records). Especially, analysis is used to perform compliance with security policies and audit, forensics, security incident response.

In order to meet the analysis role, following activities are required. First, detected event should be analyzed whether actual incident or not. Therefore, detailed information including IP address, port and packet information is required. Second, if it is an actual incident, countermeasures for incident response should be prepared.

Response. Incident response is a more professional terms. Incident response is a set of procedures for recovering abnormal to regular state of the system.

In order to meet the response role, following activities are required. First, damaged system should be isolated from the network to avoid further damage and gain more detailed information such as sample of malicious code. Second, IP address which performs malicious acts should be blocked by network security system such as firewalls. Third, malicious code should be removed from the system by using anti-virus. Finally, feedback about incident response results should be performed. This information can be used in the analysis of a similar incident.

As described above, incident handling consists of very sophisticated information security activities, and can be implemented in various ways considering the environmental characteristics. Fig.3 shows four-tier structure for incident handling:

- *Tier-1*: Information assets are main target for cyber attacks and security activities. All information assets have vulnerabilities.
- *Tier-2*: Cyber attacks are performed by exploiting vulnerabilities. Various kinds of cyber attacks can be performed by using single vulnerability.
- *Tier-3*: Security mechanisms mitigate the damage from cyber attacks on information assets. Various kinds of security solutions and techniques can be used.
- *Tier-4*: Incident handling controls cyber threat to information assets by using security mechanisms.

In this paper, we propose a practical incident handling methodology and framework focused on bot response. Especially, we choose one possible path.

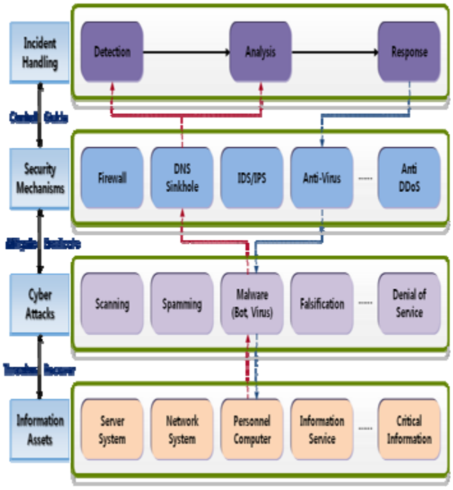


Fig. 3. Four-tier structure for incident handling

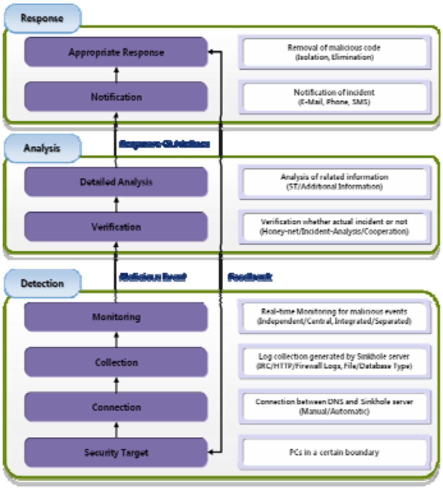

Response
Appropriate Response
Notification
Removal of malicious code (Isolation, Elimination)
Notification of incident (E-Mail, Phone, SMS)
Response Guidelines
Analysis
Detailed Analysis
Verification
Analysis of related information (ST/Additional Information)
Verification whether actual incident or not (Honey-net/Incident-Analysis/Cooperation)
Malicious Event
Feedback
Detection
Monitoring
Collection
Connection
Security Target
Real-time Monitoring for malicious events (Independent/Central, Integrated/Separated)
Log collection generated by Sinkhole server (IRC/HTTP/Firewall Logs, File/Database Type)
Connection between DNS and Sinkhole server (Manual/Automatic)
PCs in a certain boundary

Fig. 4. Overview of proposed IHM

3 Incident Handling Methodology Focused on Bot Response

Fig.4 presents an overview of proposed incident handling methodology (IHM).

3.1 Phase 1: Detection

Security Target. The security target (ST) is the PCs which are infected by bot. Because our approach is based on technical characteristics of DNS-Sinkhole that bypass C&C server connection by zombies to Sinkhole server based on DNS query. As a result, we concentrate on the monitoring for malicious activities by zombies.

The boundary of the ST is variable. Small scale of PCs within a certain organization can be the ST, and also large scale of PCs using a particular internet service provider (ISP) can be the ST. However, ST should be carefully determined by considering the efficiency of security activities. If the ST has large scale, security monitoring and control have a consistency. If the ST has small scale, faster analysis and incident response are possible.

Connection. In our approach, connection is made between DNS server and Sinkhole server. Sinkhole server must be able to perform C&C server. In order to enable this, general C&C server environment should be implemented such as internet relay chat (IRC) server, HTTP web server which do not perform malicious activities. Furthermore, it generates basic data for detection and analysis. Therefore, Sinkhole server is the most important component in proposed IHM.

Connection type can be divided as follows:

- *Manual type*: It means manually setting by DNS server administrator based on script language. It is effective in case of large scale of organization such as ISP.

- *Automatic type*: It means automatic setting by agent based on programming language. It is effective in case of small scale of organization which lacks of DNS and security expertise. In addition, it can provide scalability and convenience.

Collection. Various types of log files generated by Sinkhole server should be collected. In general, IRC and HTTP servers are used as C&C server. Therefore, we have focused on collecting IRC and HTTP logs.

Log collection can be achieved as follows:

- *IRC logs*: It can be collected from IRC daemon (IRCd) which is installed on Sinkhole server. In general, IRCd provides limited information such as connection time, IP address, channel name and nickname.
- *HTTP logs*: It can be collected from HTTP daemon (HTTPd) which is installed on Sinkhole server. However, HTTPd provides limited information.
- *Firewall logs*: Firewall logs can be collected which is provided specific operating system such as Linux and Windows. It can provide more detailed information. In particular, packet information can be helpful for future analysis.

In addition, collection type can be divided as follows:

- *File type*: Typically, daemons and firewalls create a log in the form of text files (*.txt). It is very inefficient in real-time monitoring and analysis to use text files.
- *Database type*: It means that text files should be stored in the Database. In this case, meaningful data in text files can be extracted and saved. Therefore, more efficient monitoring and analysis are available.

Monitoring. Security monitoring is a major activity for fast and accurate incident response. In particular, real-time monitoring and visibility are very important. In our approach, monitoring is based on information collected from Sinkhole server.

Sinkhole server can be located in various ways as shown in Fig.5 (red-circles are possible locations). Therefore, monitoring types can be selected depending on the location of Sinkhole server. Security target, connection and collection are also affected because monitoring is not an isolated activity.

- *Independent type*: It means that individual organization builds Sinkhole server and performs security monitoring. It is useful when organization operate private DNS. However, appropriate number of security personnel should be prepared.
- *Central type*: It means that security or network professional organization builds Sinkhole server and performs security monitoring. In particular, ISP operating public DNS or central IRT can perform this role. They can provide comprehensive incident response and analysis trends associated with cyber threats.

In addition, monitoring methods can be divided as follows:

- *Integrated type*: In general, network based security monitoring is performed by using IDS or TMS. Therefore, IDS/TMS and Sinkhole server based monitoring can be integrated. If you add simple rule (or signature) that detects any connection to Sinkhole IP address, IDS/TMS can perform the monitoring.
- *Separated type*: Separate application specialized for bot monitoring could be used. It should be able to display various information collected by Sinkhole server.

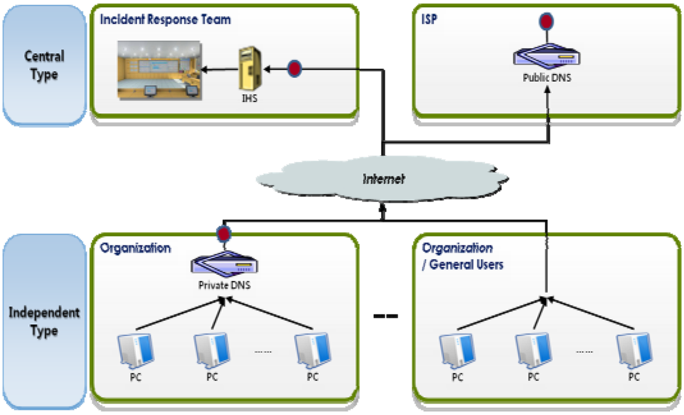


Fig. 5. Monitoring types depending on location of Sinkhole server

3.2 Phase 2: Analysis

Verification. Bot detection is highly dependent on the reliability of blacklists including actual C&C domain names. If blacklists have a high reliability, all PCs trying to connect to Sinkhole server are zombies (actual incident). However, it is very difficult because hackers change C&C frequently and botnets technology is evolving rapidly. Thus, blacklists are managed at the national level in Korea.

In general, blacklists can be obtained as follows:

- *Honey-net based*: Honey-net can be used to analysis of cyber threats trends and malicious code. Blacklists can be effectively managed based on analysis results. However, large scale of organization such as ISP and central IRT can perform this.
- *Incident-analysis based*: If actual incident occurs, blacklist can be made through analysis of malicious code. However, central IRT which has abilities to carry out comprehensive security activities can perform this.
- *Cooperation based*: Sharing of blacklists managed at the national level is also possible. If central IRTs such as national IRT, large scale of IRTs cooperate with each other, the blacklist can have high accuracy and reliability.

IRC based bots detection has very high accuracy, but other types of bots detection accuracy is relatively low. Thus, additional analysis is required to verify whether the actual incident.

Detailed Analysis. Detailed analysis provides baseline data for actual incident response. Especially, following information can be considered:

- *ST information*: Information about ST should be analyzed to future incident response. It includes user, system and similar incident history information.
- *Additional information*: More detailed information should be analyzed in order to establish countermeasures. In particular, correlation analysis can be considered. It can be achieved through the analysis of logs on information security systems such as firewalls, IDS, IPS. In addition, packet level analysis may be required to identify specific malicious actions.

3.3 Phase 3: Incident Response

Notification. Incident response should be carried out on the ST occurring actual incident. Thus, incident information analyzed through detailed analysis activity must be notified to administrator or actual user of ST. They will perform the appropriate response activities based on notified information. Especially, building and updating on network of emergency contacts should be considered additionally.

In general, the following methods can be used for notification: e-mail, phone call, short message service (SMS).

Appropriate Response. Finally, appropriate countermeasures should be applied to resolve the problem on ST. In general, the following methods can be used for incident response:

- *Isolation*: It means that ST should be isolated from the network. Zombies may perform additional malicious activities such as DDoS attack, key logging, critical data transmission to the hacker. Therefore, ST must be disconnected from the internet by using firewall/router or unplugging the network cable.
- *Elimination*: It means that malicious code installed by bot should be eliminated. It is the fundamental solution as well as final goal of IHM. In general, malicious codes can be eliminated by using anti-virus solutions.

4 Case Study

We design and implement the incident handling framework (IHF) based on proposed IHM. In particular, it can show the effectiveness of our approach, and introduce how to apply proposed IHM in actual security field.

4.1 Incident Handling Framework

Fig.6 presents the IHF. We named this framework ‘DNS-Sinkhole Service’ because it aims to provide comprehensive security service including detection, analysis and response.

Table 1 shows correlation between proposed IHM and IHF. In particular, we are trying to select optimized type of detailed activities which are focused on central IRT.

IHF has the main features as follows.

Large Scale of ST. IHF has been designed with a focus on science and technology fields in Korea. Thus, we chose one ISP and a number of organizations. In particular, ISP is the KREONET (Korea Research Environment Open Network) which is specialized in science and technology research. Also, selected organizations are composed of about 40 national institutes. As a result, approximately 60 DNS servers including one public DNS and more than 40,000 PCs have been connected to IHF.

Connection Agent. We have been developed connection agent considering convenience for automatic DNS setting and connection to Sinkhole server as well as rapidity for updating blacklists. In particular, it is a very general-purpose agent by analyzing of DNS operating environment. Also, it was implemented by using Java language which can be used on all types of OS.

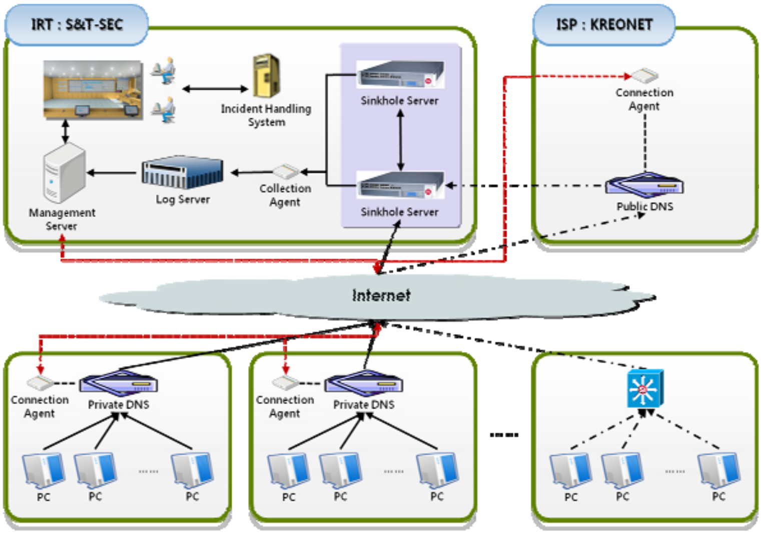


Fig. 6. IHF based on proposed methodology

Table 1. Correlation between IHM and IHF

IHM		IHF
Roles	Activities	
Detection	Security Target	ISP + Multiple Organizations
	Connection	Automatic Type
	Collection	Database Type (IRC, HTTP Logs)
	Monitoring	Central Type (Separated Type)
Analysis	Verification	Cooperation + Incident Analysis based
	Detailed Analysis	ST + Additional Information
Response	Notification	Phone + E-Mail
	Appropriate Response	Isolation + Elimination (Anti-Virus)

Fig.7 presents the structure of connection agent. It consists of four modules and each module's main functions are as follows:

- *Analysis module*: It collects and analyzes the information about OS and DNS application such as bind version and directory structure.
- *Installation module*: It sets up the zone and makes a connection to blacklist file.
- *Configuration module*: It sets up the automatic execution for agent by using scheduling program.
- *Management module*: It sets up the automatic downloading and adaptation of the blacklist file from central IRT. In particular, it performs this job once a day.

Collection Agent. We have been developed collection agent enabling real-time collection of incident related information to the database. In particular, we focused on log files from IRC daemon and Windows firewall. Both are creating logs in the form of text files. Therefore, collection agent extracts the necessary information by parsing the data in text files.

In addition, it was implemented as an application type enabling real-time monitoring by using collected information. Fig.8 presents some screen shots of the collection agent.

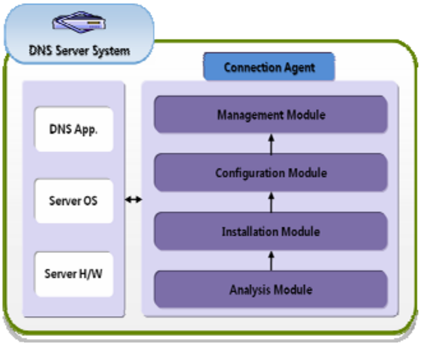


Fig. 7. Structure of connection agent

Fig. 8. Some screen shots of collection agent

Management Server. We perform comprehensive incident handling activities based on network and information security systems by using the incident handling system. In addition, we cooperate with other central IRT such as NCSC, KISA. Therefore, we constantly create and update the blacklists about C&C, then transfer it to connection agent by using management server.

We are operating website for supporting real-time monitoring by security administrators of individual organizations, and checking the incident information by personnel users of individual organizations and ISP. Especially, this website is very useful enabling interaction among the central IRT, security administrator and general users in the organizations.

4.2 Incident Response Process

We perform actual incident response activities divided into two kinds as follows:

- *Indirect response*: Once zombies are connected to Sinkhole server, they do not perform the additional malicious actions. Especially, it is very difficult to direct respond to entire PCs connected to ISP. Thus, we notify zombie lists to ISP.
- *Direct response*: The fundamental solution is elimination of malicious code. We perform direct response to approximately 40 organizations. They are managed from detection to incident response results by central IRT.

Fig.9 presents our incident response process based on IHF. Especially, it is focused on direct response.

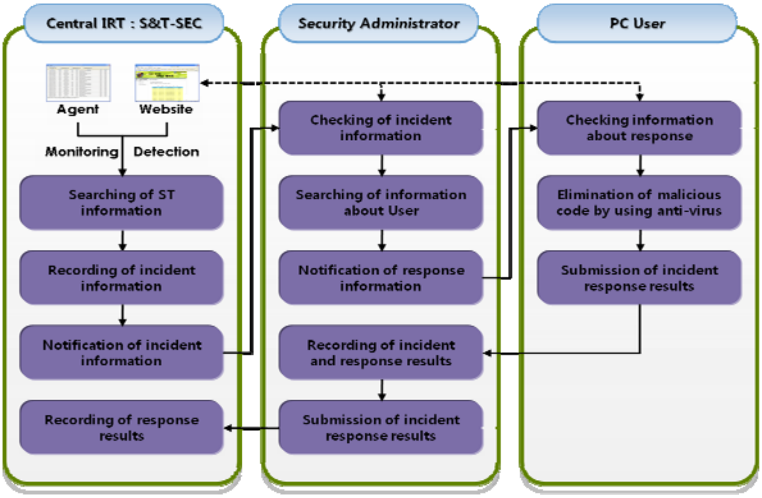

Central IRT : S&T-SEC
Security Administrator
PC User
Agent
Website
Monitoring
Detection
Searching of ST information
Recording of incident information
Notification of incident information
Recording of response results
Checking of incident information
Searching of information about User
Notification of response information
Recording of incident and response results
Submission of incident response results
Checking information about response
Elimination of malicious code by using anti-virus
Submission of incident response results

Fig. 9. Incident response process based on IHF

4.3 Operation Results of IHF

We began the operation of IHF in January 2010. Table 2 shows operation results of the last seven months.

Table 2. Operation results of IHF

Types	Results	Description
Detection	3,310	Zombie PCs infected with a malicious code
Elimination	158	Complete removal of malicious code by anti-virus
Blacklists	2,971	Domain names of C&C server analyzed by S&T-SEC

The followings are analyzed through operating IHF.

- Total number of connection attempts to Sinkhole server is 456,678. It means that zombies are trying constantly connection to C&C until it succeeds. The average number of zombie attempts to connection is about 140. In particular, zombies are trying to change the port number and protocol by regular patterns.
- Zombies based on IRC protocol were only 3. The rest of zombies were based on HTTP protocol. It means that IRC server is not typical C&C anymore. In particular, C&C server has been transformed into a variety of forms such as social network service (SNS: Twitter, Facebook).
- Elimination rates for malicious code by using anti-virus solution are only 5%. It is very difficult to force elimination of malicious code. Especially, it is impossible to direct access and control to PCs because the legal problems may occur. In addition, development of new or variant bots is exceeding the speed of pattern analysis and update for anti-virus.

5 Conclusion

In this paper, we propose a practical methodology and framework for comprehensive incident handling based on DNS-Sinkhole technique. The IHM is methodology which can cover the entire incident handling life-cycle including detection, analysis and response. The IHF is framework for actual incident handling activities based on IHM. In particular, IHF has been successfully operated for bot response by S&T-SEC which is central IRT. Therefore, it is possible to perform immediately bot response by adapting proposed framework.

In addition, proposed IHM and IHF have the scalability and flexibility although it focused on the zombies and based on specific technique. For example, other security solutions or techniques can be applied to respond to different types of cyber attacks on proposed IHM. In addition, IHF can be reduced suitable for small scale of STs.

References

1. Ianelli, N., Hackworth, A.: Botnet as a vehicle for online crime. CERT. Request for Comments (RFC) 1700 (December 2005)
2. Vogt, R., Aycock, J.: Attack of the 50 Foot Botnet. Dept. of Computer Science, University of Calgary, TR 2006-840-33 (2006)
3. Dagon, D., Gu, G., Lee, C., Lee, W.: A taxonomy of botnet structures. In: Proceedings of the 23 Annual Computer Security Applications Conference, ACSAC 2007 (2007)
4. Bailey, M., Cooke, E., Jahanian, F., Xu, Y., Karir, M.: A Survey of Botnet Technology and Defenses. In: Proceedings of Cybersecurity Applications & Technology Conference For Homeland Security (CATCH), pp. 299–304 (2009)
5. Kim, Y.-B., Youm, H.-Y.: A New Bot Disinfection Method Based on DNS Sinkhole. Journal of KIISC 18(6A), 107–114 (2008)
6. Kim, Y.-B., Lee, D.-R., Choi, J.-S., Youm, H.-Y.: Preventing Botnet Damage Technique and It's Effect using Bot DNS Sinkhole. Journal of KISS(C): Computing Practices 15(1), 47–55 (2009)
7. Korea Internet & Security Agency, A Strategy and Policy Planning for DDoS Response, KISA homepage (2010)
8. Scarfone, K., Grance, T., Masone, K.: Computer Security Incident Handling Guide, NIST SP 800-61-Revision1 (2008)
9. Bejtlich, R.: TAO of Network Security Monitoring: Beyond Intrusion Detection. Addison-Wesley, Reading (2005)

Real Time Watermarking and Encryption Method for DMB Contents Protection

Won-Hee Kim and Jong-Nam Kim

Div. of Electronic Computer and Telecommunication Engineering, Pukyong University
{whkim07,jongnam}@pknu.ac.kr

Abstract. Watermarking and Encryption are commonly technique for data protection. DMB contents are utilized widely without protection procedure. To solve the problems, we propose real-time watermarking and encryption method for DMB contents protection in this paper. To implement watermarking, we hide key information on a redundant space of program association table (PAT) and program mapped table (PMT) of T-DMB stream and hidden parts is encrypted by a stream encryption cipher. We implements encryption method without additional information in digital multimedia broadcasting (DMB) contents using AES (advanced encryption standard) encryption algorithm. In experimental result, when we implemented a play control on PMP which have built-in DSP device to get 100Mhz processing speed, almost had not a time delay and the hiding information in T-DMB stream was possible to do a play control. Additionally, we confirmed that the saved contents in a PMP were not played in other devices without decryption key. The proposed methods can be useful in real-time applications requiring contents protection service such as video on demand, IPTV and digital TV.

Keywords: Watermakring, Encryption, DMB, Contents Protection, Real Time.

1 Introduction

Most of all, portable DMB devices are possible to storing DMB contents, because of it has an internal hard-disk drive or flash memory. There devices, which is able to storing multimedia contents, are possible to an illegal distribution without agreement of the contents producers. The illegal distribution of multimedia contents must be forbid for the growth of multimedia contents distribution. The research and implementation of the contents protection technique for illegal multimedia distribution protection is progressing but, some works for protection of an illegal distribution in portable DMB devices like PMP are little. A contents protection technique for an illegal distribution protection of multimedia contents is the watermarking and the encryption. A watermarking is not pre-blocking but post-blocking about access of users and an encryption is a pre-blocking technique, which can access just authorized users, against an illegal distribution of multimedia contents.

The encryption methods for the digital video contents protection encrypt whole bit-stream domain of a digital video and a part of bit-stream or encryption progress [1-2].

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 493–501, 2010.

The farmer is possible to a strong encryption, but has high calculation complexity. The letter isn't possible to a strong encryption than the first, but has lower complexity.

In watermarking, when we classify it according to a location of hiding information, it is divided into three parts which are a spatial domain, a transform domain and bit-stream domain [3-4]. A watermark hiding of spatial domain has a lower complexity and a lower robustness, a transform domain is possible to a strong hiding of information but has higher computational complexity and a bit-stream domain has a lower complexity and a higher robustness.

In this paper, we propose a real time bit-stream watermarking method for an illegal distribution blocking of DMB contents during store contents in the internal hard-disk drive and implement on PMP. Proposed method hides encrypted play control information on redundancy parts of DMB bit-stream.

2 Related Works

There are lots of kinds of video copyright protection system. They include watermarking system, digital right management (DRM) and image encryption. Watermarking system is hiding copyright information in digital contents without decreasing image quality [3]. It is the surest way for the copyright protection, but inappropriate for blocking unauthorized users because it is not difficult for unauthorized users to access videos and then use digital contents illegally. DRM system is more effective than watermarking system in terms of conditional access. It is effectively blocking unauthorized users. DRM system prevents illegal distribution and duplication from unauthorized users.

Also, these systems permit users which use digital contents in legitimate, but protects illegal users [4]. DRM system needs network infra to share the certification key. That is difficult to use some DMB devices not having networking function. In addition, it is very expensive for implementation because DRM systems need the expensive computer systems and networking elements.

Image encryption is to distribute the partially or fully encrypted images [1]. There are many kinds of encrypt algorithms - DES, AES, SEED, and so on. DES algorithm was invented in 1977. However, this encryption algorithm is broken in 1997, so it is sure that it is not robust. To replace DES, AES algorithm was developed [2].

AES algorithm is symmetric key block algorithm and block size is 128bit key size is variable from 128bit to 250bit. Encryption and decryption speed in digital signal processor is a positive point faster than SEED and RC6 [5].

Encrypted images cannot be replayed without certification key. It is an efficient way to block unauthorized users. Image encryption technique has a weak point. That is, there is no way to block the distribution of decrypted images from legitimate users. The way we suggest is to operate encryption and decryption on a PMP and impossible to distribute the decrypted contents. Existing method of encryption technique encrypts the I-frame or P-frame in video data. However, it was not yet implemented in embedded devices (PMP or another DMB device) [6-8].

There are two kinds of DMB, in which one is Terrestrial DMB (T-DMB) and the other is Satellite DMB (S-DMB). These two DMB systems use the same video compression algorithm. For the compression, H.264/MPEG-4 Part 10 AVC technology is used. Our implemented system is based on T-DMB.

3 The Proposed Methods

In the section, we introduce proposed watermarking method and explain implantation on PMP. An implementation of a real time bit-stream watermarking on a portable DMB device is difficult to an implementation using a conventional method. In case of DMB, it is consisted of a packet unit of TS and is possible to a real time watermarking when we do watermarking in especially small part of TS packets [9]. Proposed method encrypts hiding watermark information for a play control using by Dragon stream cipher and then hides play control information in redundancy data which isn't a active video data in TS packets. Fig. 1 show a structure of DMB transport stream (TS). TS have a length of 188 bytes and are consisted of header and payload. Headers of TS packet have various information which is sync information, transport error information, PID information, transport priority information, adaptation field control information and so on. We use PID information to find program allocation table (PAT), program mapped table (PMT) and packetized elementary stream (PES) in header information of TS packets.

Fig. 1. To a watermarking DMB bit-stream, we have to find PAT or PMT packet in TS packets. If PID value of TS packet is '0', it indicates PAT. Fig. 2 shows PAT structure. PMT is presented one program map PID in PAT.

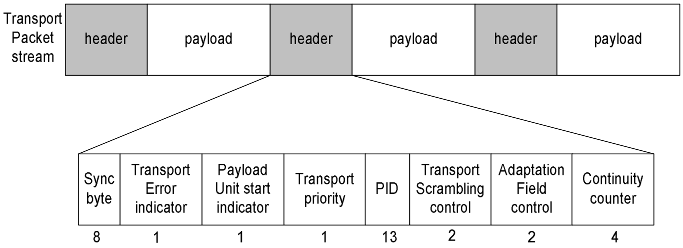


Fig. 1. A structure of DMB transport stream

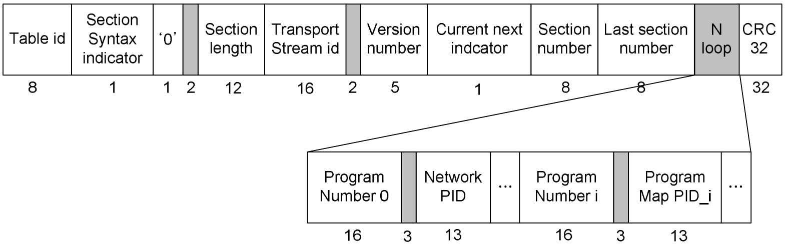


Fig. 2. PAT structure

Program map PID is included information to know where program map table (PMT) in TS. PMT is payload of TS such as PAT and include PID for elementary stream location in TS, the structure of PMT show to Fig. 3. IOD_descriptor and elementary stream ID is included in PMT, so we can find PID which includes packetized elementary stream (PES) location in TS without additional descriptors. PES is compressed video data of DMB contents.

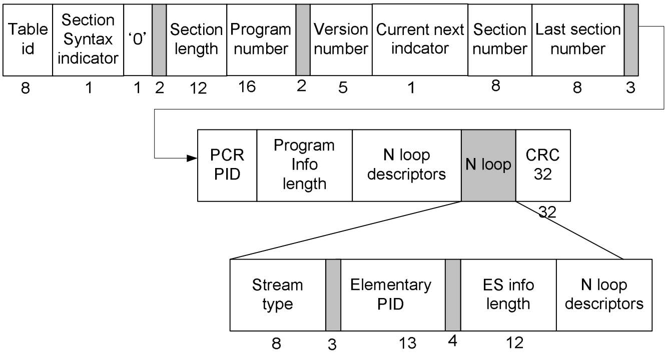


Fig. 3. PMT structure

In our implementation method, we must find padding data space in PAT and PMT. For finding padding data, we find PID '0' with TS parsing in receiving DMB stream and then find PAT on next TS packet and find PMT part in TS with 'Program map ID' in PAT. And we find value of 'Section length' in PAT and PMT.

We used TVUS 900 PMP from Homecast, Fig. 4 shows DMB processing procedure in PMP. Our implemented system embeds a watermark for play control on video stream data which is a result of TS parsing module of DMB and extracts a hidden watermark on video stream data.

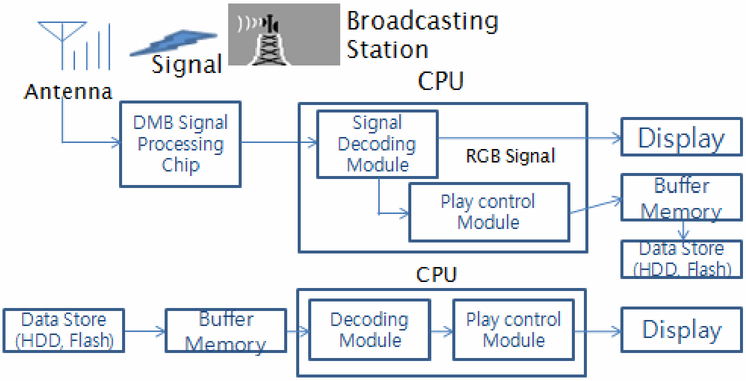


Fig. 4. Proposed PMP system structure

A TVUS of Homecast using DM320 main processor is related to internally four threads for a DMB play. Each thread is BSI, BSO, MAF and GUI. GUI handles about it in case of manipulating button on PMP and informs start of run to BSI, BSO and MAF using a start event when open the media file. BSI transfer frame information of video according to GUI event order to MAF and MAF store decoded video data using DSP in shared buffer. Saved video frame format is YCbCr(4:2:2). BSO convert YCbCr to RGB format to play stored video frame.

In DMB stream a watermarking on PMP, first of all we stored one TS packet (188 bytes size) using TSD module of BSI thread in the buffer, and then extract PAT and PMT in TS packet. PAT and PMT are indicated to 'Section Length". It indicates padding information in packet.

In this paper, we find watermarking parts for play control in DMB bit-stream, we analyzed PAT packet. PAT header has 'Selection Length' which indicates PAT data length in TS packet. If a length of PAT data sufficiently small, remaining parts except PAT data length is filled with padding data which haven't any data information. Padding data hasn't any information about a video data and header, we do hiding the play control information without any change of bit-stream format and size.

In DMB stream encryption in PMP, first of all is stored in the buffer one TS packet (188 bytes size) using TSD module of BSI thread, and then extract PAT and PMT in TS packet. PAT and PMT information is used to find PID of PES in TS packet. If payload of TS packet is PES, then we select encryption parts according to the data in the buffer. First 4 bytes of the buffer which is stored PES packet is TES header, next 18 bytes is PES header and after 22th byte in the buffer is video data parts.

In this paper, to find DMB encryption parts in PES packet check PID which know a kind of payload in PES packet. If PES payloads are video data, PID is '0x50' or '0x113'. PID indicate a video data, next step check an adaptation field value, adaptation field is payload start location in PES packet, if adaptation field value is '00b' or '10b' then do not exist payload in PES packet. Start point of payload in the buffer is IDR slice, IDR slice is included I-frame information. From data of first macro block to end data of macro block in IDR slice is encrypted by AES algorithm and then stored in PMP. If it is not IDR slice or first macro block in IDR slice then we do not encryption. Decryption procedure is loaded encryption video data in PMP hard disk using internal program of PMP, and then find encrypted payloads using the same method of encryption procedure. We decrypt encrypted macro blocks using by AES decryption algorithm.

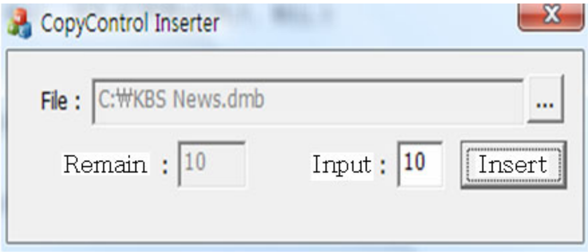
4 Experimental Results

Used to implement the system is PMP and PC. Used PMP is TVus900 model. It made by HOMECAST Corporation, Korea. Software development tool is Microsoft's eMbedded Visual C++ 4.0 and Microsoft's Platform Builder 5.0. We ported PMP with debugging board offered from HOMECAST Corporation. Also, firmware source was offered from there.

For performance assessment, we processed the experiment as following conditions. For play encryption DMB video and decryption on PC. Specification of PC is Intel Pentium-4 2.8GHz CPU, 1GByte RAM and Microsoft Windows XP SP3. To play DMB contents in PC, we can use OnTimeTek Corporation's DMBO Filter and Gretech Corporation's GOM Player. Fig. 5 shows PMP used by this implementation.

Fig. 6 shows a play control configuration program. It is program for a play control test and this program is activated on PC. In figure, 'File' is sequence to insert the play control information and 'Remain' indicates a remaining number for DMB contents playing and 'Insert' is a number which is inserted for playing of DMB contents.

Fig. 5. Homecast™ TVUS 900 PMP



CopyControl Inserter

File : C:₩KBS News.dmb

Remain : 10

Input : 10

Fig. 6. A play control configuration program

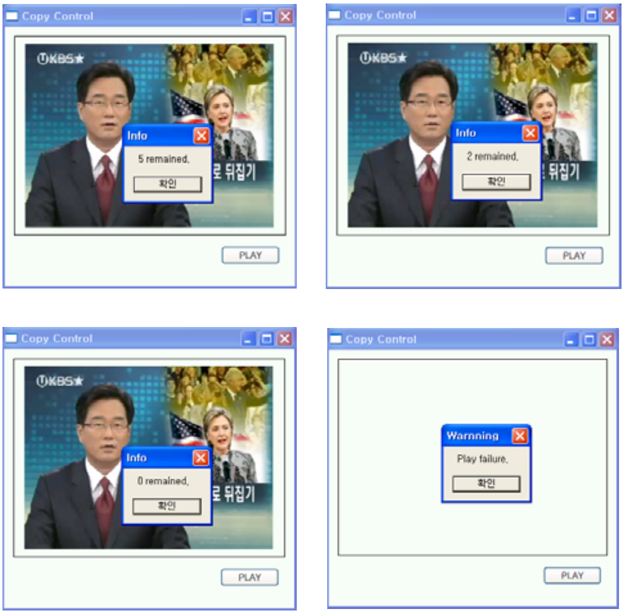
Fig. 7 shows a play control program of DMB contents. This program is possible to a play control using hiding information in padding data of TS packets. DMB contents of Figure 8 is inserted play control information (in this test, we use 5).

Encryption test program is designed like PMP's encryption and decryption module using Microsoft Visual Studio 8.0's MFC code.

Fig. 8 is DMB contents encryption and decryption result using Test program. Fig. 8(a) is original content. Fig. 8(b) is encryption result of 8(a). Fig. 8(c) is decryption result of 8(b).

We used DMB contents for encryption recording ten from PMP. Used contents for test are the size of MPEG4/AVC, 320x240 that is Korean standard, and one I-frame and 29 P-frame make up one GOP (Group of Picture).

Fig. 9 is the experimental result of DMB encryption module in PMP. Fig. 9(a) is the played content in PC, not encrypted content. It is not applied to protection (encryption), we can see clearly play. Fig. 9(b) is the played Fig. 9(a) content in PMP. However, this content is not permitted to be played in PMP. Therefore, we cannot see any image but it is distorted and destroyed. Fig. 9(c) is played in PMP which is encrypted in the PMP. That content is authorized in PMP. It can be played in the PMP. Fig. 9(d) is played in PC, encrypted content. It is not played in PC that is unauthorized. Therefore, this content must not be played in PC.



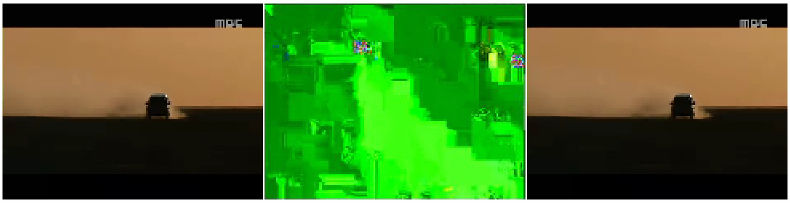
5 remained.

2 remained.

0 remained.

Play failure.

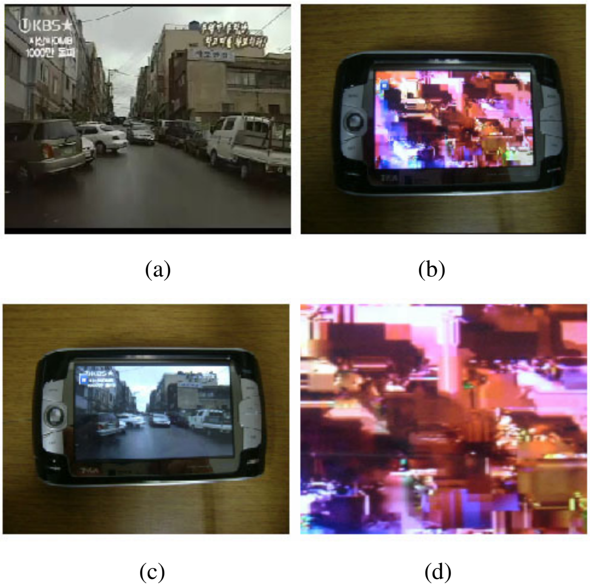
Fig. 7. A play control shot of DMB contents



(a) (b) (c)

Fig. 8. Result of encryption and decryption on PC

After encrypting and recording content in test, we transfer recording files from PMP to PC and replay them in PC. In consequence, all contents did not be played normally. Also, we replay decrypted recording files in PMP applied to encryption module. Then all contents did not be played normally and output destroyed in PMP.



(a) (b)

(c) (d)

Fig. 9. Result of experiment

5 Conclusions

In this paper, we proposed real-time watermarking and encryption method for DMB contents protection. Suggested real-time watermarking hides a play control information in padding data of TS packets. . Suggested real-time encryption system encrypts only I-frame using AES algorithm. In consequence, it is not allowed to view the contents without decryption module and certification key. It is possible to use our suggested algorithm's solidity into practice. In experimental result, when we implemented a play control on PMP which have built-in DSP device to get 100Mhz processing speed, almost had not a time delay and the hiding information in T-DMB stream was possible to do a play control. Additionally, we confirmed that the saved contents in a PMP were not played in other devices without decryption key. The proposed methods can be useful in real-time applications requiring contents protection service such as video on demand, IPTV and digital TV.

Acknowledgments. This research was financially supported by MEST and KOTEF through the Human Resource Training Project for Regional Innovation, and supported by LEADER.

References

1. Wu, M., Mao, Y.: Communication-friendly encryption of multimedia. In: IEEE Workshop on Multimedia Signal Processing, pp. 292–295 (December 2002)
2. Doomun, M.R., Soyjaudah, K.S., Bundhoo, D.: Energy consumption and computational analysis of rijndael-AES. In: 3rd IEEE/IFIP International Conference in Central Asia on Internet, pp. 1–6 (September 2007)

3. Cox, I., Miller, M., Bloom, J.: Digital watermarking. Press of Morgan Kaufmann, San Francisco (2001)
4. Nishimoto, Y., Baba, A., Kurioka, T., Namba, S.: A digital rights management system for digital broadcasting based on home servers. IEEE Transaction on Broadcasting 52(2), 167–172 (2006)
5. Daemen, J., Rijmen, V.: The Design of Rijndael: AES - The Advanced Encryption Standard. Springer, Heidelberg (2002)
6. Qiao, L., Nahrstedt, K.: A new algorithm for MPEG video encryption. In: Proceeding of International Conference on Imaging Science, Systems, and Technology, pp. 21–29 (July 1997)
7. Liu, J., Zou, L., Xie, C., Huang, H.: A two-way selective encryption algorithm for MPEG video. In: International Workshop on Networking, Architecture, and Storages, p. 5 (August 2006)
8. Zheng, L., Xue, L.: Motion vector encryption in multimedia streaming. In: Proceedings. 10th International Multimedia Modelling Conference, pp. 64–71 (January 2004)
9. T-DMB white paper Press of Electronic and Telecommunication Research Institute, Korea (December 2006)

A Study on Data Transmission Performance of Sensor Networks for Livestock Feedlot

Ji-woong Lee, Hyun-joong Kang, Jeong-hwan Hwang,
Meong-hun Lee, and Hyun Yoe*

School of Information and Communications Engineering, Sunchon National Univ.,
315 Maegok-dong, Sunchon-Si, Jeonnam 540-742, Korea
leejiwoong, hjkang, jhwang, leemh777, yhyun@sunchon.ac.kr

Abstract. Recent development of versatile small size and multifunctional wireless sensor nodes enables the research on various applications to improve human life with rich information and automation. In this study, we virtually simulated the efficiency of the livestock feedlot sensor network by utilizing the propagation model, as one of the different methods of collecting data regarding livestock in feedlot is researched. As indicated in the conclusion of the study, the differences in terms of the methods make a difference to the amount of dropped data packets. It is believed that the observations made in this study could prompt the development of more effective methods of collecting data regarding livestock in feedlots by adopting additional devices or alternative routing manner.

1 Introduction

In feedlots, where a large number of livestock is raised in a group, the chances that the livestock will be exposed to disease and that such disease will spread rapidly [1][2]. Accordingly, many studies have been conducted with the aim of predicting the potential diseases of the livestock, using a sensor-based network, which involves the measurement of the temperature, symptom of disease, activity status, etc. of the livestock being raised in feedlots [3][4][5][6]. Most feedlots employ the method of breeding a certain number of livestock in a group by installing hedges of lumbers on a spacious pasture and setting up partitions between them. Although a large number of the livestock raisers are more conscious of possible diseases, it is very difficult to transmit measured data from a Wireless Sensor Networks (WSN), where the sensor is driven by low power due to the differences in propagation performance on account of the position of the sensor attached, interrupting by the livestock being raised, etc. Blocking elements like these will certainly require a more effective method of data transmission from a livestock-feedlot-based sensor network. The chief aim of this study is to try to compose the measurement of data for a large number of livestock, which is difficult to apply, in virtual reality, via simulation. Much research has been conducted on the routing efficiency in a WSN on account of the channel condition of

[*] Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 502–509, 2010.

the physical layer (PHY) according to such blocking elements as indoor furniture [7][8]. As far as is known, however, no research has yet been conducted on the subject of livestock. Although some simulations have been done, in a similar form, through individual link paths, on man according to sensor's location [9], schemes other than those proposed in the foregoing studies are required because livestock, unlike man, do not act intelligently, and their behavior is different from that of men.

The rest of this paper is organized as follows. Section 2 reviews propagation models. In Section 3, we present the conducted simulation environment in detail. Section 4 describes the simulation results. Finally, we conclude this paper in Section 5.

2 Related Research (Propagation Models)

In this section, we explain the current propagation models for simulating the physical layer for researching the relation between physical layer condition and data exchanging performance. The propagation model has huge influence on performance of WSN. A model depends on various parameters like the distance between sensor nodes. But others expressed as random functions and constant factors [10].

2.1 Free Space (FS) Model

The free space model is the simplest propagation model. It only presumes the direct path between transmitter *t* and receiver *r*. The receiving power P_r depends on the transmitted power P_t, the gain of the receiver and transmitter antenna (G_t, G_r) the wavelength λ, the distance d between both nodes and a system loss coefficient *L*. All parameters, but the distance *d*, are system wide constant parameters. While a simulation runs, the receiving power P_r only changes with the distance between sender and receiver. As both receiving parameters RX_{Thresh} and CS_{Thresh} are also constant during a simulation, receiving nodes must be inside a perfect circle. Otherwise, they are unable to receive data properly.

$$P_{r,FS} = \frac{P_t \cdot G_t \cdot G_r \cdot \lambda^2}{(4\pi \cdot d)^2 \cdot L}$$

2.2 Two Ray Ground (TRG) Model

The TRG model is an improved version of the FS model. It assumes the direct ray between sender and receiver and also considers the ground reflection. As with the FS model, both nodes are assumed to be in LOS. The heights of both antennas over the ground are present with h_t and h_r and are constant during a simulation. Up to the crossover distance $4_{Threshtrh}=\pi\ h_t\ h_r\ \lambda d$, the TRG model is equal to the FS model. Beyond this distance, the ground reflection destructively intervenes with the direct ray and further reduces the field strength. The receiving signal strength is then inverse proportional to d^4. Just like the FS model, TRG has only the distance between sender and receiver as variable parameter *H*.

$$P_{r,TR} = \begin{cases} P_{r,FS} & d < d_{Thresh} \\ \dfrac{P_t \cdot G_t \cdot G_r \cdot h_t^2 \cdot h_r^2}{d^4 \cdot L} & d \geq d_{Thresh} \end{cases}$$

2.3 Shadowing Model

For both FS and TRG models, the sender-receiver distance is the only variable parameter during a simulation. This forms a circular coverage around a sending node and has a cutting range limit. Beyond this range, further reception is impossible. To introduce random events, the shadowing model uses a random variable *X*. The shadowing model needs a reference distance d0 to calculate the average received FS signal strength $P_{r,FS}(d_0)$. The path loss exponent β is influenced by the simulated environment and is constant throughout simulations. Values vary between two (free space) and six (indoor, non-line-of-sight). *X* is normal distributed with an average of zero and a standard deviation σ (called shadow deviation). Again it is non-variable and reasonable values vary between three (factory, LOS) and twelve (outside of buildings). Values for β and σ were practically determined.

$$P_{r,SH} = P_{r,FS}(d_0)\left(\frac{d}{d_0}\right)^{-\beta} \cdot 10^X$$

$$X(x): \{x \in [-\infty, \infty] \,|\, P(x) = N(0, \sigma^2)\}$$

3 Simulation

3.1 Simulation Scenarios

The scenario that was studied in this research consisted of 24 mobile sensor nodes in an area 100m × 100m big, and the sink node for collecting the sensing data was placed at the center of the simulating environment. Based on such an environment, simulation was done on two occasions.

In scenario 1, the livestock moved about freely in the feedlot, and the mobile nodes (attached to the livestock) moved at a slow rate and transmitted data periodically.

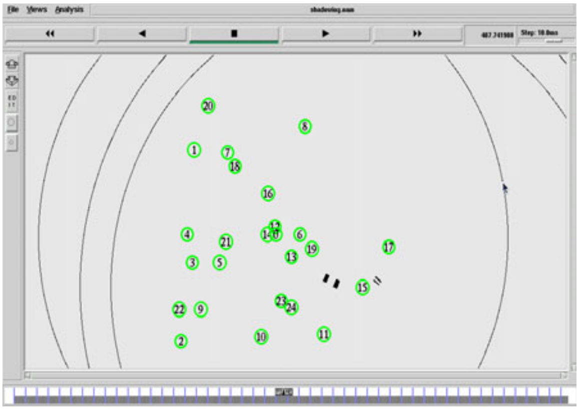


Fig. 1. Scenario 1, Sensor nodes move freely

In the other case, in scenario 2, using the behavioral characteristics of livestock, it was assumed that the sink node was installed at the feeder from which the livestock feed. Also in the second scenario, the mobile nodes got together at the feeder in the center, while the sink node, a collecting node, collected data from each of the animals in the system.

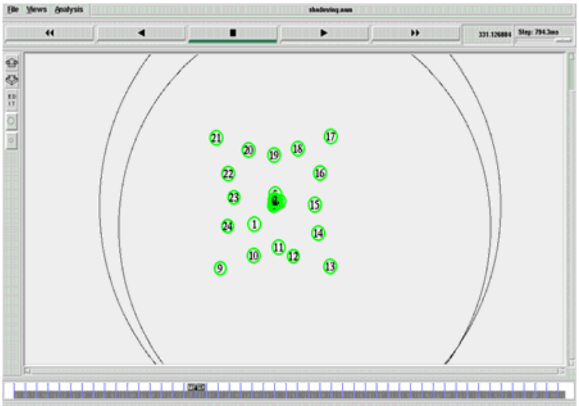


Fig. 2. Scenario 2, Sensor nodes move to the sink

3.2 Route Setup and Data Delivery Process

To determine the route path between sensor nodes, initially, sensor nodes search adjacent nodes within the transmission range. Then request for the path information toward a sink node or destination node. From this information, a sensor node forward data to the destination node through other sensor nodes using multi-hop manner. In case, if there is no adjacent node or no information from nodes within the transmission range, a node periodically send a discover message or Query message.

3.3 Propagation Model

The NS-2 simulator [11], which is most commonly used at present, was used as the simulator in this simulation. In the first assumption (scenario 1), in the system where data were collected while the livestock moved around freely in the feedlot, it was decided that the livestock had different channel conditions according to their distance from the central sink node. Although in NS-2, the propagation model by livestock within the demand has not yet been defined, since there is such a propagation model as a shadowing model, the blocking by livestock was composed in such a way that, according to distance, different shadowing models were applied among different sensor node links.

3.4 Simulation Environment

It was assumed that there was no moving of the livestock to another place, no exception caused by death, etc., and that the sensor node had ample initial energy that could not be used up within the period of the simulation.

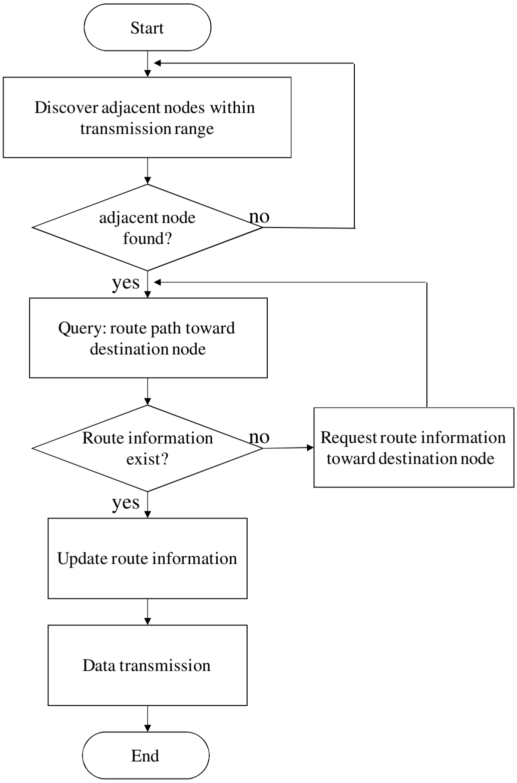

Start
Discover adjacent nodes within transmission range
adjacent node found?
no
yes
Query: route path toward destination node
Route information exist?
no
Request route information toward destination node
yes
Update route information
Data transmission
End

Fig. 3. Route setup and data delivery process

For the simulation, the basic parameters of NS-2 were used, together with a Lucent WaveLAN radio interface at 914 MHz. As for the antenna, it was assumed that a unity-gain omni-directional antenna was placed at the center of the node, with a height of 1.5m. $RX_{Threshold}$ of Wireless Phy was adjusted to receive data with 95% probability from a 20m distance, which allowed different scope of coverage to be made among different links.

In the simulation, which was carried out for 1,000sec, all the nodes, except the Sink node, transmitted information towards the Sink node in a 5sec cycle, with a transmitted data size of 512 bytes. The initial energy was placed at 10,000 joules, which was deemed sufficient for the entire duration of the simulation, with 0.660 joules used for data transmission and 0.395 for reception. In both the mentioned simulations, it was supposed that a mobile node for transmitting biological data was attached to the livestock, and that all the nodes moved at a low rate of 0.1 m/sec. Furthermore, in scenario 1, where the nodes moved around freely, each node was made to have different direction and moving distance values, while in scenario 2, where the nodes moved at the same rate while they were gathering towards the center.

The MAC layer was not considered, and routing was made using Destination Sequence Distance Vector (DSDV) [12], Ad-hoc On-demand Distance Vector (AODV) [13], and Dynamic Source Routing (DSR) [14], which are adhoc-based typical proactive and reactive routing protocols. DSDV, a representative proactive routing protocol, is based on the Bellman-Ford routing system used in wired networks, and each node always keeps routing information regarding all the other nodes in the routing table. AODC, a very light routing protocol for ad-hoc networks, is the routing protocol that is used to make routing possible while using less memory. DSR is an on-demand routing protocol that is based on a routing source. Mobile nodes maintain a route cache, including the source routes known to a node.

4 Simulation Results

Figure 4 and figure 5 represent dropped data packets from each routing protocol. In the case of scenario 1, AODV and DSDV universally kept the dropped packets uniformly, while DSR revealed a severe shift span in the amount of dropped packets through scenario 1 and 2. It assumed that as the movement of sensor node is unpredictable, streamlined data transmission is quite difficult so the dropped data packets occur during the simulation continuously.

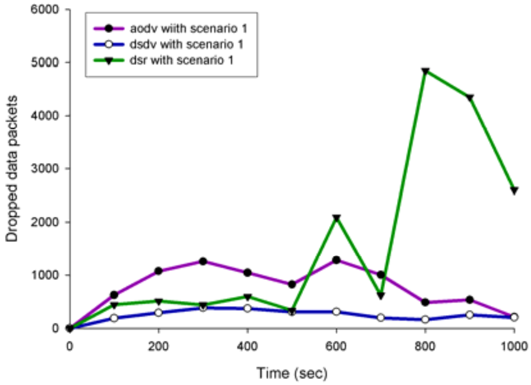


Fig. 4. Dropped packets with Scenario 1

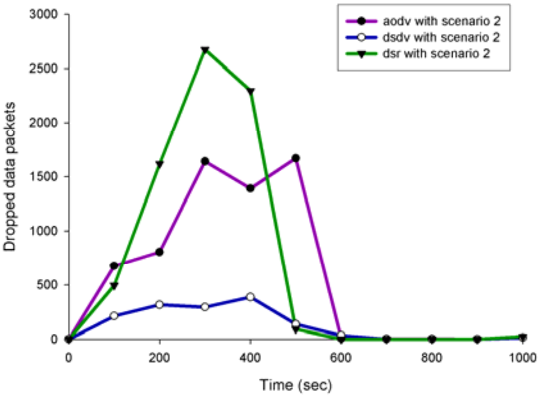


Fig. 5. Dropped packets with Scenario 2

In the case of scenario 2, at the initial state, dropped data packets are severely increased and it was shown to decrease gradually, starting from around 300 sec, and to hardly appear at 600 sec and over. It assumed that as sensor nodes gather to the center of simulation area, link path toward the sink node is constantly changes. After 600 seconds and later, data dropping is considerably decreased it is because the node movement is declined and variation of position of sensor node is almost predictable.

5 Conclusion

In this study, difference in efficiency between routing protocols in the aspect of dropping packets was researched. For this purpose, the research was conducted with two different scenarios aimed to examine the influence on routing protocols related to the data transmission of sensor nodes caused by free movement of livestock. With the results from simulation, there is a certain difference between two scenarios. It can thus be concluded that diverse methodical attempts like global positioning system(GPS) or relative position anticipate method to reduce the dropping packets caused by movement are needed for the collection of data of freely moving nodes regarding livestock in feedlots and an efficiently streamlined routing protocol design is needed.

Acknowledgements

"This research was supported by the MKE(Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)"(NIPA-2010-(C1090-1021-0009)).

References

[1] Smith, R.A.: Impact of disease on feedlot performance: a review. Journal of Animal Science (1998), http://animal-science.org

[2] Jim, G.K., Booker, C.W., Ribble, C.S., Guichon, P.T., Thorlakson, B.E.: A field investigation of the economic impact of respiratory disease in feedlot calves. Can. Vet. J. 34, 668–673 (1993)

[3] Wang, N., Zhang, N., Wang, M.: Wireless sensors in agriculture and food industry—Recent development and future perspective. Computers and Electronics in Agriculture 50, 1–14 (2005)

[4] Nagl, L., Schmitz, R., Warren, S., Hildreth, T.S., Erickson, H., Andresen, D.: Wearable sensor system for wireless state-of-health determination in livestock. In: Proceedings of the 25th IEEE EMBS Conference, Cancun, Mexico, September 17–21 (2003)

[5] Radenkovic, M., Wietrzyk, B.: Wireless mobile ad-hoc sensor networks for very large scale livestock monitoring. In: Proc. of ASWN, Berlin, Germany, pp. 47–58 (2006)

[6] Maatje, K., de Mol, R.M., Rossing, W.: Cow status monitoring (health and oestrus) using detection sensors. Computers and Electronics in Agriculture 16(3), 245–254 (1997)

[7] Kreuzgruber, P., Unterberger, P., Gahleitner, R.: A ray splitting model for indoor radio propagation associated with complex geometries. In: 43rd IEEE Veh. Technol. Conf. Secaucus, NJ, pp. 227–230 (May 1993)

[8] Obayashi, S., Zander, J.: A body-shadowing model for indoor radio communication environments. IEEE Transactions Antennas and Propagation, 920–927 (1998)

[9] Habaebi, M.H., Abduljali, E., Ragaibi, T., Rhuma, N.: Effect of sensor specific body location on wireless network routing performance. Electronics Letters, 40–42 (January 3, 2008)

[10] Gruber, I., Knauf, O., Li, H.: Performance of Ad Hoc Routing Protocols in Urban Environments. In: Proceedings of European Wireless (2004)

[11] Fall, K., Varadhan, K.: The ns Manual (formerly ns Notes and Documentation). The VINT Project, A Collaboration between researches at UC Berkeley, LBL, USC/ISI and Xerox PARC

[12] Perkins, C., Bhagwat, P.: Highly dynamic destination-sequenced distance-vector (dsdv) routing for mobile computers. In: Proceedings of ACM SIGCOMM (1994)

[13] Perkins, C., Royer, E.: Ad hoc On demand Distance Vector Routing. In: Proc. of 2nd IEEE Workshop on Mobile Computing Systems and Applications (February 1999)

[14] Johnson, D., Maltz, D.: Dynamic source routing in ad hoc wireless networks. In: Imielinski, T., Korth, H. (eds.) Mobile Computing, The Netherlands, pp. 153–181. Kluwer Academic Publishers, Dordrecht (1996)

Development of an MPEG-4 Scene Converter for Rich Media Services in Mobile Environments

Hee-Sun Kim and Ilgun Ko

Department of Multimedia Engineering, Andong National University
388 Songchun-dong, Andong-si, Gyungbook 760-749 Korea
hskim@andong.ac.kr, ilgunko@gmail.com

Abstract. MPEG-4 BIFS is a major rich media standard that has been selected as the standard technology for data broadcasts and interactive contents on DMB. BIFS defines various audio/visual nodes and two- and three-dimensional graphic nodes, as well s the interactions of many users, but it is not appropriate for the mobile environment with limited performance. Therefore, this study proposes a method of converting MPEG-4 BIFS to an MPEG-4 LASeR format that is appropriate to the mobile environment, to service rich media in a mobile environment. First, the scene structure and nodes of MPEG-4 BIFS and LASeR were comparatively analyzed. It is not difficult to convert BIFS to LASeR for most 2D nodes. Since LASeR does not support 3D, however, it is difficult to convert 3D nodes to LASeR. To solve this problem, this study extracted the co-ordinate characteristic values of BIFS 3D nodes and converts them to 2D coordinates to express LASeR as SVG. To test this method of conversion, a system that converts BIFS to LASeR was created. This allowed the provision of diverse rich media services, including 2D and 3D services, from a mobile device.

Keywords: MPEG-4, LASeR, Scene converter.

1 Introduction

MPEG-4 defines BIFS, a language that structures audio/visual objects and their scenes, for two-way conversation-type services[1]. BIFS is a major rich media format that has been selected as the standard for conversation contents and data broadcasting through DMB. BIFS can express rich media well through various nodes and abundant user interaction, but it is not appropriate to the mobile environment that has limited performance. With the fundamental start of digital broadcasting, it has become very important to service rich media in various environments, especially mobile devices. Therefore, there is a need to study how to service rich media such as BIFS in the mobile environment.

Therefore, this study proposes a method of converting the existing MPEG-4 BIFS to the MPEG-4 LASeR (Lightweight Application Scene Representation: ISO/IEC 14496-29)[2,3] format that is appropriate to the mobile environment, to service rich media in a mobile environment. LASeR uses the 2D graphic language of W3C,

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 510–517, 2010.

SVG[4], to express graphic animation and handle light resources, and is structured by limited scene technology nodes that are appropriate to the mobile environment.

Studies related to MPEG-4 scene conversion focused on creating MPEG-4 BIFS as XMT[5] or converting XMT to VRML or SMIL. One study on converting XMT to BIFS focused on the authoring tool developed by IBM[6]. This is a 2D scene writing tool that creates XMT alpha and omega, and includes functions that create XMT and convert XMT-A to BIFS. There has been a study on converting MPEG-4 contents to LASeR to play them on mobile devices[7]. This study converted only 2D nodes that can convert BIFS to LASeR. This is because LASeR uses SVG, the 2D graphic language, and it is possible to convert 2D-graphics-related nodes of BIFS to LASeR, but 3D-related nodes are not supported and are difficult to convert. With the demand for contents such as 3D navigation or 3D games growing in the mobile environment, however, the expression of LASeR-based 3D nodes has become important.

Therefore, this study presents a method of converting 3D nodes to LASeR. This study presented a system of converting XMT files, the textual format of BIFS, to LASeR, and designed and implemented the system. The conversion system parses the XMT files that the user entered to create a DOM tree on the memory. 2D nodes with functions and characteristics similar to those of XMT and LASeR are converted using the XSLT[8] style sheet. 3D geometrical nodes, however, must go through a process of converting the node value to a value that can be handled through a LASeR node. Next, a substitute node that can be mapped as a LASeR standard SVG node on a DOM tree is added, and the value is revised. This conversion method can service rich media, including 3D nodes, through mobile devices.

The second part of this study comparatively analyzes MPEG-4 BIFS and LASeR, and the third part explains the method of converting BIFS to LASeR. Part four shows the results of the materialization of the system that converted the BIFS text format XMT to LASeR, and part five presents the conclusion.

2 Comparison of BIFS and LASeR

The standard plan of BIFS is limited because it is a PC-based standard and demands many resources for the mobile environment. MBEG-4 is therefore an alternative to BIFS, and MOEG-4 part20 LASeR was set as the standard. The characteristics of BIFS and LASeR are shown in Table 1.

BIFS uses VRML as its base technology and can express 2D and 3D scenes, but the standard is complex. On the other hand, LASeR is a base technology that uses SVG and through which only 2D graphic expressions are possible. Moreover, it has a simple standard. In addition, it includes technology and videos where the document itself is not downloaded but scene renewal and conversion are made into binary contents and streamed, and technology that expresses freely together audio and graphic data and logically streams them in multiple transmissions. Therefore, LASeR enables dynamic rich media services that support abundant scene expressions and efficient compression in a mobile environment.

Table 1. Comparison of the characteristics of BIFS and LASeR

	BIFS	LASeR
Standard	MPEG-4 Part1 System	MPEG-4 Part20 LASeR
Device	PC	Mobile
Base Technology	VRML	SVG
Scene Presentation	2D/3D	2D
Binary Encoding	Support	Support
Interactive	Support	Support
Streaming	Support	Support

3 Method of Conversion of BIFS to LASeR

3.1 Conversion of BIFS to LASeR Using the XSLT Conversion Method

XSLT is used to convert XMT files, the text format of BIFS, to LASeR. XSLT is a markup language that explains the process of converting XML-based documents to XML documents with different structures. The node conversion rules for the conversion of BIFS to LASeR are created as an XSL file for the conversion of BIFS documents to LASeR documents. Table 2 shows the nodes that can be converted immediately through one-to-one counteraction because they have similar functions and characteristics.

Table 2. Nodes that can be converted one-to-one from BIFS to LASeR

BIFS nodes	LASeR nodes
Geometry Nodes(Rectangle, Curve2D, Circle, shape, appearance …)	svg : rect , svg : circle svg : path
Text, FontStyle	svg : text , svg : tspan
Hyperlink nodes(inline, anchor)	svg : a , svg : foreightObject
Transform2D node	svg : g
OrderedGroup, Layer2D	svg : g
WorldInfo node	svg : svg
ImageTexture, PixelTexture MovieTexture nodes	svg : image, svg : video
Interpolator, TimeSensor nodes	lsr : animate
Touch Sensor	lsr : trigger node
Conditional	lsr : script
AudioClip, AudioSource nodes	svg : audio
Switch nodes	svg : g
Valuator node	Unmapped
PointSet2D, CompositeTexture2D TextureCoordinate nodes	Unmapped
TempCap, QuantizationParameter	Unmapped

During the sconversion of XSLT from BIFS to LASeR, since the coordinate systems of BIFS and LASeR differ, the conversion of the coordinate values must be considered. The BIFS 2D node coordinates values are accordingly changed to the LASeR coordinates. It is difficult to convert 3D nodes to XSLT through LASeR, though. Table 3 shows BIFS 3D nodes that cannot easily be converted to LASeR. SVG, the scene technology language of LASeR, cannot express 3D coordinate values because it shows 2D graphics. The 3D geometrical objects that are used in BIFS such as boxes, cylinders, cones, and spheres, and parent nodes such as transform and material nodes, are difficult to convert. Therefore, this study presented a method of converting the nodes that are difficult to convert, as shown in Table 3, and materialized the conversions.

Table 3. Nodes that are not easily converted from BIFS to LASeR

Class	Node Name	2/3D	Types/Function
General-use	Coordinate	3D	Attribute
General-use	Material	3D	Attribute
General-use	Transform	3D	Group
Geometry	Box	3D	Geometry
Geometry	Sphere	3D	Geometry
Geometry	Cylinder	3D	Geometry
Geometry	Cone	3D	Geometry
Geometry	IndexedLineSet	3D	Geometry
Geometry	IndexedFaceSet	3D	Geometry

3.2 Method of Converting BIFS 3D Graphic Nodes to LASeR

This study separately investigated BIFS 3D nodes and proposed a method of adding nodes to allow their handling as LASeR nodes, so as to change the BIFS 3D nodes to LASeR nodes. Figure 1 shows how BIFS 3D geometric nodes can be expressed as LASeR nodes.

First, as shown in Figure 1 (1), the XMT-A file is parsed and a DOM-Tree is created on the memory. This DOM-Tree, which is shown in Figure 1 (2), finds 3D nodes and extracts attribute values. For example, the box object of XMT has size attribute, and the attribute is used to determine the reference coordinates. The value extracted in Figure 1 (3) is used to create reference coordinates that structure a 3D geometric object. Boxes have size attribute and create reference coordinates that show eight points of the box object, with the starting point as the standard. The coordinates, rotation, and scaling of the XMT nodes are stored in the parent node that is the transform node, so that as shown in Figure 1 (4), the attributes of the transform node are extracted and the coordinates of the object are rotated, scaled, and calculated. For example, the location movement information, the rotate information, and the scale information of the XMT transform node are used to move, rotate, and scaled the reference coordinates in Figure 1 (3) and to find the location where the object will be expressed. Next, the 3D coordinates are projected and converted into 2D coordinates, as shown in Figure 1 (5), and the 2D coordinates are again converted.

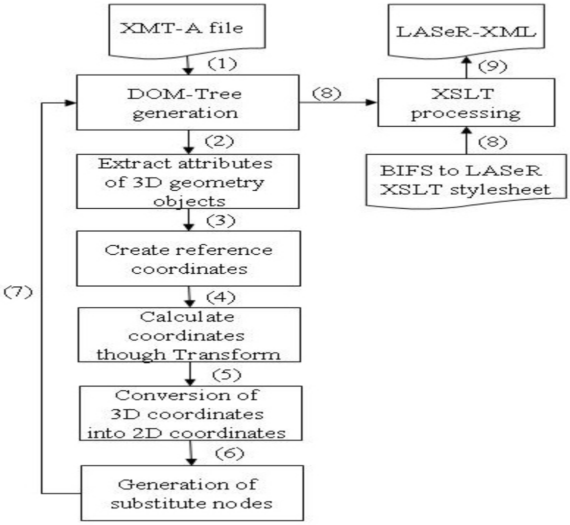

XMT-A file
(1)
DOM-Tree generation
(2)
Extract attributes of 3D geometry objects
(3)
Create reference coordinates
(4)
Calculate coordinates though Transform
(5)
Conversion of 3D coordinates into 2D coordinates
(6)
Generation of substitute nodes
(7)
(8)
LASeR-XML
(9)
XSLT processing
(8)
BIFS to LASeR XSLT stylesheet

Fig. 1. Method of converting BIFS 2D geometric nodes to LASeR nodes

BIFS defines the center of the scene as the starting point, but LASeR defines the top left as the starting point. Therefore, the coordinates must be converted. Figures 1 (6) and (7) add a substitute node to express the 3D node of BIFS as the LASeR expression on the XMT DOM-tree. Instead of the 3D geometric objects of BIFS, i.e., the box, sphere, cylinder, and cone, alternative nodes with 2D coordinate values, such as a 2D box, a 2D sphere, a 2D cylinder, and a 2D cone are added. In the case of the indexed face set and the indexed line set, the coordinate-related values are converted into 2D coordinate values. Next, in Figures 1 (8) and (9), the XSLT conversion is performed to convert XMT to LASeR.

The box is expressed as svg:polygon; the cone, as svg:circle and svg:polyline; the sphere, as svg:ellipse; and the cylinder, as svg:polygon and svg:ellipse. The indexed line and the indexed face are calculated when there is a transformation, and after the 3D coordinates are converted into 2D, the values are changed and the indexed line 2S and the indexed face set 2D are mapped in the same way on LASeR.

4 Implementation of the Conversion System

The conversion system was implemented in a Windows XP environment to Java using the Eclipse 3.4 development platform. Using the conversion system presented in this study, XMT-A files were converted to LASeR. Figure 2 shows the results of the use of the mp4Box of an XMT-A file GPAC that includes 3D geometrical objects to create mp4 files, and the results of playing them on IMI-3D, the IBM MPEG-4 contents player. Figure 2 shows the 3D nodes of BIFS--the box, cylinder, sphere, cone, and indexed face set. Some of the XMT-A files in Figure 2 are shown in Table 4.

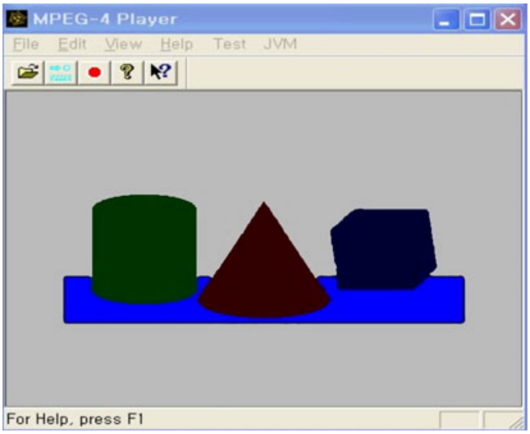


Fig. 2. Results of playing XMT-A, which includes the 3D nodes

Table 4. Parts of the XMT-A file in Figure 2

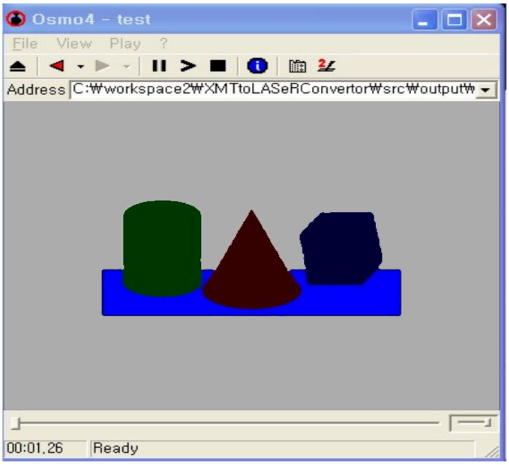
```
<?xml version="1.0" encoding="UTF-8"?>
<XMT-A xmlns="urn:mpeg:mpeg4:xmta:schema:2002"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="urn:mpeg:mpeg4:xmta:schema:2002 xmt-a.xsd">
 <Header>
     ... omission ...
 </Header>
 <Body>
 <Replace>
  <Scene>
 <Group>
     <children>
     <Transform rotation="1 0 0 0.349" translation="-180 0 0" scale="1 1 1">
      <children>
       <Shape>
        <appearance>
         <Appearance>
          <material>
           <Material diffuseColor="0 1 0"/>
          </material>
         </Appearance>
        </appearance>
        <geometry>
         <Cylinder height="160" radius="80"/>
        </geometry>
   ... omission ...
 </Body>
 </XMT-A>
```

The XMT-A file in Table 4 was converted using the conversion system suggested in this study. Table 5 shows the LASeR-XML file, which is the result of the conversion of Table 4. The 3D nodes of XMT in Table 4, i.e., the box, cylinder, sphere, cone, and indexed face set, are expressed in Table 5 as g and the polygon and ellipse of LASeR.

Table 5. LASeR-XML file that is the conversion result of Table 4

```
<?xml version="1.0" encoding="UTF-8"?>
<saf:SAFSession xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:lsr="urn:
mpeg:mpeg4:LASeR:2005" xmlns:saf="urn:mpeg:mpeg4:SAF:2005" xmlns:ev=
"http://www.w3.org/2001/xml-events" xmlns="http://www.w3.org/2000/svg">
<saf:sceneHeader>
<lsr:LASeRHeader profile="full" colorComponentBits="8"/>
</saf:sceneHeader>
<saf:sceneUnit>
<lsr:NewScene>
<svg id="root-node" width="480" height="320" viewport-fill="rgb(0,0,0)">
<g>
<polygon points="96.0,221.7 384.0,221.7 384.0,174.7 96.0,174.7 " fill="rgb(0, 0, 255)"/>
<polygon points="294.2,190.1 315.6,166.5 368.2,166.5 346.8,190.0" fill="rgb(0, 0, 55)"/>
<polygon points="284.6,137.5 306.0,114.0 358.6,114.0 337.2,137.5" fill="rgb(0, 0, 55)"/>
<polygon points="346.8,190.0 368.2,166.5 358.6,114.0 337.2,137.5" fill="rgb(0, 0, 55)"/>
<polygon points="294.2,190.1 346.8,190.0 337.2,137.5 284.6,137.5" fill="rgb(0, 0, 55)"/>
<polygon points="315.6,166.5 294.2,190.1 284.6,137.5 306.0,114.0" fill="rgb(0, 0, 55)"/>
<polygon points="315.6,166.5 368.2,166.5 358.6,114.0 306.0,114.0" fill="rgb(0, 0, 55)"/>
<ellipse cx="240.0" cy="195.8" rx="48.0" ry="19.6" fill="rgb(55, 0, 0)"/>
    ... omission ...
</saf:SAFSession>
```

Figure 3 shows the results of the conversion of XMT to LASeR in Table 5, and of playing it on the GPAC Osmo4 multimedia player. The results confirmed that the XMT file that included 4F nodes used LASeR, which supports only 2D for the same method of expression.



Address C:₩workspace2₩XMTtoLASeRConvertor₩src₩output₩

00:01.26 Ready

Fig. 3. Example of playing Table 5

5 Conclusion

This study proposed a method of converting 3D-related graphic nodes of BIFS to LASeR for the inter-operation of MPEG-4 contents in a mobile environment, and materialized a conversion system using the proposed method. The problem of 3D-graphics-related node conversion, particularly the difficulty of converting BIFS to LASeR, was solved. The 3D node of BIFS, the transform, and the nodes such as the coordinate, material, indexed line set, indexed face set, box, cylinder, sphere, and cone, could be expressed as a LASeR g, polyline, polygon, circle, and ellipse. Through this conversion method, rich media, including 3D nodes, can be serviced from a mobile device.

References

1. Understanding MPEG-4: Technologies, Advantages, and Market (An MPEGIf White Paper), The MPEG Industry Forum
2. ISO/IEC 14496-20:2006(E) Information technology - Coding of audio-visual objects - Part 20: Lightweight Application Scene Representation(LASeR) and Simple Aggregation Format (SAF)
3. Duford, J.-C., Avaro, O., Concolate, C.: LASeR: the MPEG Standard for Rich Media Servicse, IEEE Multimedia
4. W3C, Scalable Vector Graphics (SVG) Tiny 1.2 Specification, `http://www.w3.org/TR/2005/WD-SVGMobile12-20050413/`
5. Kim, M, Wood, S.: XMT: MPEG-4 Textual Format for Cross- Standard. Interoperability, Overview of XMT. IBM Research (2004)
6. IBM Toolkit for MPEG-4, `http://www.alphaworks.ibm.com/tech/tk4mpeg4`
7. Shahab, Q.M.: A Study on Transcoding of MPEG-4 BIFS scenes to MPEG-4 LASeR Scenes in MPEG-21 Digital Item Adaptation Framework, School of Engineering Information Center Univ. (2006)
8. Burke, E.M.: Java and XSLT. O'Relly, New York (2001)

Dynamic Profiling for Efficiency Searching System in Distributed Computing

Chang-Woo Song[1], Tae-Gan Kim[1], Kyung-Yong Chung[2],
Kee-Wook Rim[3], and Jung-Hyun Lee[1]

[1] Dept. of Information Engineering, INHA University
[2] Dept. of Computer Information Engineering, Sangji University
[3] Dept. of Computer and Information Science, Sunmoon University
up3125@hotmail.com, taegan@empal.com, kyjung@sangji.ac.kr,
rim@sunmoon.ac.kr, jhlee@inha.ac.kr

Abstract. RFID technology that identifies objects on request of dynamic linking and tracking is composed of application components supporting information infrastructure. Despite their many advantages, existing applications, which do not consider elements related to real-time data communication among remote devices, cannot support connections among heterogeneous devices effectively. As different network devices are installed in applications separately and go through different query analysis processes, there happen the delays of monitoring or errors in data conversion. This paper proposes recommendation service that can update and reflect personalized profiles dynamically in Distributed Computing environment for integrated management of information extracted from RFID tags regardless of application. The advanced personalized module helps the service recommendation server make regular synchronization with the personalized profile. The proposed system can speed and easily extend the matching of services to user profiles and matching between user profiles or between services. Finally dynamic profiling help to reduce the development investment, improve the system's reliability, make progress in the standardization of real-time data processing in matching searching system.

Keywords: Profiling, Searching System, Distributed Computing.

1 Introduction

With the advance of information technologies and the spread of Internet use, the volume of usable information is increasing explosively. In order to deal with users' activities at home including learning, leisure, and housework, we need technologies to locate users' position and to provide services according to users' current situation. Various user positioning technologies are being proposed including GPS (Global Positioning System) for outdoor positioning, AT&T Active Bat using ultrasound and RF (Radio Frequency) signals for indoor positioning, and Smart Floor technology that locates users' position using sensors installed on the floor.

However, these positional technologies are used limitedly due to restricted environment and difficulty in commercialization. Furthermore, as each of service providers

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 518–525, 2010.

has to manage a large volume of profiles separately, they cannot diversify their services to users. Accordingly, we need to develop a method for the effective management of user profile information such as users' current situation, terminal capacity, and personal preference[1]. In applications of various areas, information systems, network systems, etc. are involved in daily life, coping actively with users' actions such as housework, production and amusement, providing intelligent services fittest for users' situation, and displaying the process and results of services[2,3,4]. In order to provide such intelligent services, we need a system that integrates and manages user-related information for choosing adequate services based on RFID technology and application-related information such as communication protocol and control methods[5]. The system models personalized data and stores user profile information such as preference, behavioral pattern and correlation together with ontology. It also analyzes the ontology model and extracts data through domain ontology concept modeling, and produces an ontology-based profiling mechanism and personalized services. Moreover, it defines user profiling, ontology domain model, and personalized semantic profile model that enables semantic inference. For the inference of domain ontology and the provision of personalized data, the proposed system as a semantic user profile manager provides users with more accurate profiling.

2 Related Work

2.1 Localization Using RFID

This Comprehensive and accurate user information is necessary for integrating many different recommendation algorithms. A generic customer profile plays a very important role in collecting position as user information in this paper. Fig. 1 describes the structure of localization method using RFID.

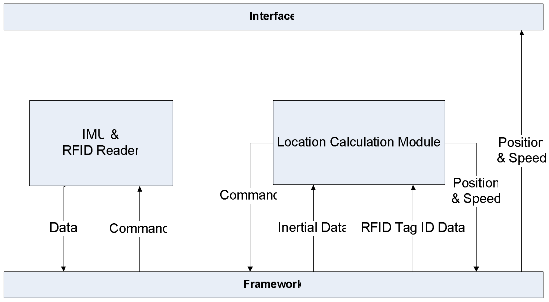


Fig. 1. Structure of Localization method using RFID

This Comprehensive and accurate user information is necessary for integrating many different recommendation algorithms. A generic customer profile plays a very important role in collecting position as user information in this paper. Fig. 1 describes the structure of localization method using RFID.

A. Inertial Based User Position Estimation

A method that estimates user position based on inertial sensors measures the present location of users by accumulating the traveling distance of users. Thus, the starting point in such traveling is required to estimate the location of users. This will accumulate errors occurred by the measurement of sensors, which are operated for long hours. Thus, this paper uses a proximity position measurement method using RFID in order to solve the accumulation of these errors and develops a position estimation method using inertial sensors based on this method. The user position determined by a position estimation manager can be used to classify domains in a voice recognition system.

B. Proximity Position Measurement

RFID is a technology that transfers information, which identifies human beings or objects, without any connections using wireless frequencies. It can be classified as passive and active types according to the use of batteries. The passive type does not include batteries and obtains required energies from wireless radio waves. Thus, it is a semi-permanent device and low price. Whereas, the active type includes batteries and shows limited service life time. However, it is able to actively collect and process data from surroundings. The objective of the use of RFID is to manage disaster, production, distribution system, and stock system. In addition, it has been used in other fields, such as path tracing and antitheft systems, based on the property of wireless data storing and transferring. The principle of proximity measurements using RFID is that the detected location in which a tag, which is attached to a mobile object, is detected by an RFID reader is determined as the present position. It is possible to estimate the traveling path and direction of mobile objects according to the time order in the detection of tags. Although it represents an advantage that is able to obtain uniformed location information, it is necessary to increase the degree of concentration for installed tags. However, an increase in the degree of concentration in tags shows a decrease in the recognition level of tags due to mutual interferences.

C. Position Initialization

The position estimation method that uses inertial sensors proposed in this paper calculates the position of users by adding the traveling distance and direction data obtained from inertial sensors based on the initial position of users to the initial position data. Therefore, it is possible to obtain and apply absolute positions that correspond to real spaces instead of relative positions by configuring the obtained initial position of users. Because new position data will be used to initialize the present position, the problem of accumulated errors can also be initialized due to the initialization of such errors.

It requires a wait period for the recognition of tags by a reader after performing an initialization command for using an RFID reader. As tags are recognized by the reader, the ID of the recognized tag can be searched in a table, which stores the position of tags. If a corresponding ID exists in the table, it can be configured as the present position value after bringing the position of the stored ID.

2.2 NaÏve Bayesian Contextual Classification

In Bayesian analysis, the effects of the prior distribution and the significance threshold selection tend to vanish as the data accumulate. Also demonstrated that our Bayesian classifiers, built by carefully modeling the specific properties of noisy data such as e and s, outperformed Support Vec-tor Machine (SVM), the most commonly used classification method (see Supplementary Material)[10].

The Bayes theory refers X as sample data, which has no classification in its class, and H can be assumed that X is included in a class of C. The possibility in which H will occur as sample data of X is generated can be determined as P(H|X) and that is called by prior probability. We need P(X), P(H), and P(X|H) to calculate P(H|X). These measurements are possible through the paper data and that is called by posterior probability.

In this paper, there is a n dimensional characteristics vector, X, determined by {context1, context2, …, contextn} in the Naive Bayesian. Also, it is assumed that there exist m context information classes determined by {class1, class2, …, classm}. The estimation in which the context information of X will be included in the class, which has the highest posterior probability, can be calculated using the Bayes theory.

Because P(X) represents constant values for all classes, it is necessary to maximize P(X|Ci)P(Ci) only. If the prior probability of a class is not determined, only P(X|Ci) is to be considered. The P(X|Ci) can be calculated using the independent assumption of Naive Bayesian as (1), and the calculation result of X can be classified as the context information class that has the highest posterior probability.

$$\begin{aligned} P(X \mid C_i) &= P(x_1, x_2, ..., x_n \mid C_i)P(C_i) \\ &= P(x_1 \mid C_i)P(x_2 \mid C_i) \cdots P(x_n \mid C_i)P(C_i) \\ &= P(C_i)\prod_{k=1}^{n} P(x_k \mid C_i) \end{aligned} \tag{1}$$

The Bayes theory refers X as sample data, which has no classification in its class, and H can be assumed that X is included in a class of C. The possibility in which H will occur as sample data of X is generated can be determined as *P(H|X)* and that is called by prior probability. We need *P(X)*, *P(H)*, and *P(X|H)* to calculate *P(H|X)*. These measurements are possible through the paper data and that is called by posterior probability.

In this paper, there is a n dimensional characteristics vector, X, determined by {context1, context2, …, contextn} in the Naive Bayesian. Also, it is assumed that there exist m context information classes determined by {class1, class2, …, classm}. The estimation in which the context information of X will be included in the class, which has the highest posterior probability, can be calculated using the Bayes theory.

Because *P(X)* represents constant values for all classes, it is necessary to maximize *P(X|Ci)P(Ci)* only. If the prior probability of a class is not determined, only *P(X|Ci)* is to be considered. The *P(X|Ci)* can be calculated using the independent assumption of Naive Bayesian as (1), and the calculation result of X can be classified as the context information class that has the highest posterior probability.

3 Dynamic Profiling

3.1 Advanced Personalized Profile

As the system proposed in this paper generates personalized profiles containing users' preference and various lifestyles and uses the recommendation service, it updates based on past history and currently available services. It builds user profiles by interpreting different types of data from RFID, making inferences from the data, and generating information. Then, combining the profiles with the services information database, the system recommends services through filtering.

Fig. 2 illustrates the architecture of dynamic profiling for recommendation service in automation system is largely composed of profile collector, profile aggregator, and collector resoner. When an event takes place, information generated by the profile collector is analyzed, integrated, and delivered to the collector resolver. The collector resoner converts the analyzed data into data include contextual information understandable to the computer through RDF ontology object databases and the RDF inference engine, and transfers the data to the profile accumulator. Then, the profile aggregator synchronizes with the server's personalized module and generates a personalized profile.

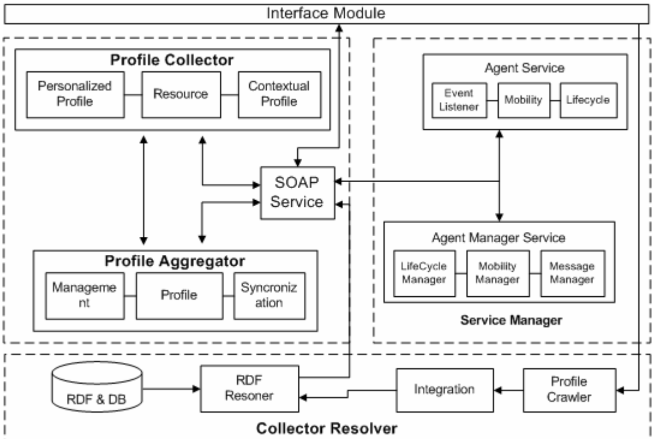
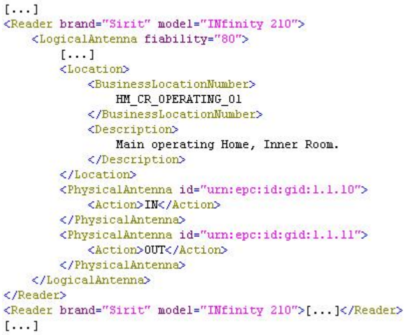


Fig. 2. Architecture of Dynamic Profiling for Recommendation Service

3.2 Information Translation

In order to standardize information demanded by each application, the database processing system converts profiles into RDF. RDF is a knowledge expression language that extends elements for context information modeling, and is used for profile matching. RFID reader information entered in XML as shown Fig 3. goes through the style resetting logic and is converted to RDF. RFID tag information is also processed by the database processing system and converted to RDF[6]. The style resetting logic is designed in XSL. The database server update module has functions such as adding triples, updating accurate values, and integrating identical expressions among triples.



```
[...]
<Reader brand="Sirit" model="INfinity 210">
    <LogicalAntenna fiability="80">
        [...]
        <Location>
            <BusinessLocationNumber>
                HM_CR_OPERATING_01
            </BusinessLocationNumber>
            <Description>
                Main operating Home, Inner Room.
            </Description>
        </Location>
        <PhysicalAntenna id="urn:epc:id:gid:1.1.10">
            <Action>IN</Action>
        </PhysicalAntenna>
        <PhysicalAntenna id="urn:epc:id:gid:1.1.11">
            <Action>OUT</Action>
        </PhysicalAntenna>
    </LogicalAntenna>
</Reader>
<Reader brand="Sirit" model="INfinity 210">[...]</Reader>
[...]
```

Fig. 3. XML Data of RFID Information

In order to process information expressed in RDF, the RDF data query language (RDQL) is used. RDQL provides a method of using the state of the declaration section, which must be processed under conditions given to the application developer. In this paper, reader and tag information are processed by the database processing system using RDQL[7,8].

3.3 Profile Mapping Model

The use of semantic language RDF can provide users with special types by defining rules. (2) shows how RFID tag information should be expressed for its delivery when the smart system of User is connected.

$$\begin{aligned} User \equiv \exists\, hasInterest.(&Mltimedia \\ &\sqcap \exists\, hasGenre.Movie \\ &\sqcup \exists\, hasHome.DVDPlaying \\ &\sqcup \exists\, hasLocation.LivingRoom \\ \sqcap \forall\, hasInterest.(&\neg Game \\ &\sqcup \neg \exists\, hasHome.VideoGame) \end{aligned} \tag{2}$$

One excludes uninteresting game applications and watches recommended movies on a DVD player at Living Room in home. The semantic service can provide users with special types using rules. In this paper, two types of rules are defined. ‘Offers’ is a standard method of service provision. The definition of special types between the service provider and users is extended. ‘Demands’ is defined through matching between service and profile, and is called by users’ request.

$$\begin{aligned} Profile \equiv\ & \sqcap_i \exists\, hasInterest.Interest\ _i \\ & \sqcap_j \forall\, hasInterest.(\neg DisInterest_j) \end{aligned} \tag{3}$$

As in (3), the semantic user profile expresses preference as ‘Interested’ or ‘Uninterested.’ By applying the rules of service request and provision, profile mapping is done as in (4, 5). The goal of profile mapping is to determine whether a given profile is semantically compatible with a specific service. In addition, it determines how to match. Ontology compares users’ preference with a service provider.

$$\begin{aligned} &UserInterest \equiv \exists\, hasInterest^{-1}.Profile \\ &ServiceOffer \equiv \exists\, offers^{-1}.Service \end{aligned} \tag{4}$$

$$Match \equiv UserInterest \sqcap ServiceOffer \tag{5}$$

If a user enters a specific space, the initial location is set through installed RFID tags. While the user is moving, data such as acceleration, angular velocity and tag ID are collected from IMU and transmitted. If position estimation is initiated, the user waits until the RFID tag is recognized. When the RFID tag is recognized, the RFID Position Manager checks if the tag ID is included in the list of RFIDs with predefined position. If the tag has a predefined position, the current service location is initialized with the position of the tag. Receiving data include contextual information from inertial sensors, the Location Calculation module calculates the current service location after the user has moved based on the initialize location. The Profile Manager identifies the user by the service request, and compares the semantic profile in the form of RDF containing preference and lifestyle with the current service location. Then, service matching information is transmitted according to the profile.

4 Conclusion and Future Works

We have through the paper work seen a need for improvement in system architecture with respect to system integration, dynamic profiling handling related to automation system. The advanced personalized module helps the service recommendation server make regular synchronization with the personalized profile include information actively. The proposed method can speed and easily extend the matching of services to user profiles and matching between user profiles or between services. Dynamic profiling help to reduce the development investment, improve the system's reliability, make progress in the standardization of real-time data processing in matching automation system. In order to use different forms of profile information, the system defines profiles in RDF, and the reuse of profiles in interface access and the provision of various services bring many benefits to users and produce an effect equivalent to direct selection. In addition, it was found that several profiles are linked to profile resources in profile collection.

However, the proposed framework needs to be evaluated using various context recognition services, case studies, and prototypes demanding feedback. Further work is needed with respect to stakeholder interaction and protocols for that interaction. So, we plan to conduct quantitative analysis on the implicit rating extraction mechanism.

Acknowledgments. "This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) Support program supervised by the NIPA(National IT industry Promotion Agency)" (NIPA-2010-C1090-1031-0004).

References

1. Zhou, X., Wu, S.-T., Li, Y., Xu, Y., Lau, R.Y.K., Bruza, P.D.: Utilizing Search Internet in Topic Ontology-based User Profile for Web Mining. In: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (2006)

2. Kobayashi, I., Saito, M.: A Study on Information Recommendation System that Provides Topical Information Related to User's Inquiry for Information Retrieval. In: Web Intelligence and International Agent Technology Workshops, pp. 385–388 (2006)
3. Jung, K.Y., Lee, J.H.: User Preference Mining through Hybrid Collaborative Filtering and Content-based Filtering in Recommendation System. IEICE Transaction on Information and Systems E87-D(12), 2781–2790 (2004)
4. Keenoy, K., Levene, M.: Personalization of web search. In: Mobasher, B., Anand, S.S. (eds.) ITWP 2003. LNCS (LNAI), vol. 3169, pp. 201–228. Springer, Heidelberg (2005)
5. Miller, B.N., Konstan, J.A., Miller, J.R., Konstan, J.A., Pocketlens, J.R.: PocketLens: Toward a Personal Recommender System. ACM Transactions on Information Systems (TOIS) 22(3), 437–476 (2004)
6. Kim, Y.H., Kim, B.G., Lim, H.C.: The index organizations for RDF and RDF schema. In: ICACT 2006, vol. 3, pp. 1871–1874 (2006)
7. Beckett, D.: RDF/XML Syntax Specification, W3C (2004)
8. Baader, F., Horrocks, I., Sattler, U.: Description Logics as Ontology Languages for the Semantic Web. In: Hutter, D., Stephan, W. (eds.) Mechanizing Mathematical Reasoning. LNCS (LNAI), vol. 2605, pp. 228–248. Springer, Heidelberg (2005)
9. Huang, H., Liu, C.C., Zhou, X.J.: Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. Proc. Natl. Acad. Sci., 6823–6828 (2010)

Evaluating the Intention to Use the Use Case Precedence Diagram in Software Development Projects

José Antonio Pow-Sang

Departamento de Ingeniería, Pontificia Universidad Católica del Perú,
Av. Universitaria 1801, San Miguel, Lima 32, Peru
japowsang@pucp.edu.pe

Abstract. The Use Case Precedence Diagram (UCPD) is a technique that addresses the problem of determining a software's scope and construction sequence from the developer's perspective. This paper presents a qualitative evaluation of the UCPD based on the Method Adoption Model (MAM), where the intention to use a method is determined by the users' perceptions. The results show that the intentions to use UCPD exist in undergraduate students and practitioners with at least 2 years of experience in the industry, but the relationships defined by the MAM are only confirmed with the results obtained with practitioners.

Keywords: UCPD, requirements precedence, software engineering experimentation, Method Adoption Model.

1 Introduction

Use cases technique was first proposed by Ivar Jacobson [11], and, since its inclusion in the Unified Modeling Language (UML) [16], its utilization has been greatly extended, making it a mandatory requirement for any software development project. The Use Case Precedence Diagram (UCPD) [18] is a technique based on use cases and its objective is to determine software construction sequences taking into consideration the developer's perspective in terms of ease of construction to define software requirements priorities.

According to UML, the relations that can exist between use cases are: include, extend, and generalization. In addition to the standard, UCPD proposes the inclusion of a new relation: precedence. The concept of this diagram was taken from Doug Rosenberg [20], who proposed the use of a similar diagram, specifying the relations "precedes" and "invoke" to determine user requirements. Fig. 1 shows an example of a UCPD. There are two rules to define the precedence relationship between use cases, one considers the precondition, and the other whether a use case needs the information that is registered by another use case (further details of these rules can be found in [18]).

The use cases that are on the left side of the diagram will be implemented before the ones that are on the right side. For instance, in Fig. 1, "Use case A" will be implemented before "Use case C".

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 526–537, 2010.

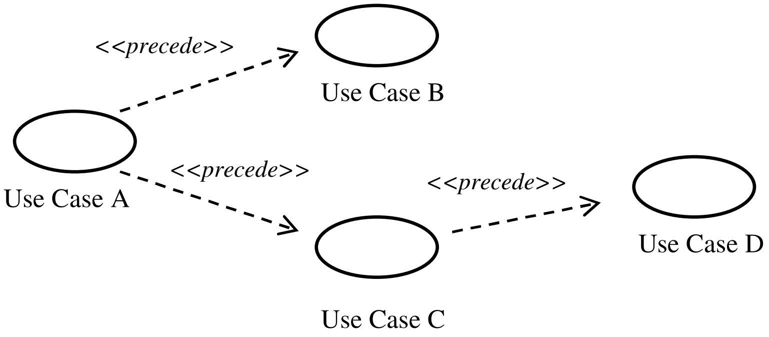


Fig. 1. Use Case Precedence Diagram

In [19], it was included the results of a controlled experiment in which UCPD is applied in case studies by practitioners and the obtained results show that UCPD has more significant advantages over the utilization of ad-hoc techniques.

Although the results obtained in the controlled experiment were satisfactory, there is a need also to assess users' response to the new procedure and their intention to use it in the future, for which reason we applied at the end of the experiment a questionnaire based on the Method Adoption Model (MAM) [14].

MAM was proposed by Moody and this model is an adaptation of the Technology Acceptance Model [9] defined by Davis. MAM explains and predicts the adoption of methods. The constructs of the MAM are the following:

- *Perceived Ease of Use:* the extent to which a person believes that using a particular method would be effort-free.
- *Perceived Usefulness:* the extent to which a person believes that a particular method will be effective in achieving the intended objectives.
- *Intention to Use:* the extent to which a person intends to use a particular method.

MAM defines that perceived usefulness is influenced by the perceived ease of use, and intention to use is defined by perceived usefulness and perceived ease of use. Many empirical studies that evaluate software methods have been carried using MAM with students and practitioners [1,7,17]. Some of them do not confirm the relationships between the constructs defined by the MAM.

The rest of the paper is organized as follows: Section 2 describes this study, Section 3 details the results obtained for the empirical study. Finally, a summary and our plans for future research will conclude our paper.

2 Description of This Study

Using the Goal/Question/Metric (GQM) template for goal-oriented software measurement [3], we defined this study as follows:

Analyze: user's responses
For the purpose of: evaluate
With respect to: intention to use UCPD

From the point of view of: the researcher
In the context of: undergraduate students and practitioners with at least 2 years of experience in software development projects, considering that the developer is free to select the sequence to construct use cases (there are no user's constraints).

Based on the MAM model, it was formulated the working hypotheses of this research which make reference to the intrinsic constructs of the model. These hypotheses are stated in the following way:

- Hypothesis 1 (H1): UCPD is perceived easy to use.
- Hypothesis 2 (H2): UCPD is perceived useful.
- Hypothesis 3 (H3): There is an intention to use UCPD in future software projects.
- Hypothesis 4 (H4): The perceived ease of use has a positive effect on the perceived usefulness of UCPD.
- Hypothesis 5 (H5): The perceived ease of use and perceived usefulness has a direct and positive effect on intention to use.

2.1 Participants

The undergraduate students who participated in this study were fourth year students of the Informatics Engineering program at the Pontificia Universidad Católica del Perú (PUCP) that were enrolled in the Spring '06 Software Engineering course. As part of the course, these undergraduate students had to develop a software using the Rational Unified Process methodology [10] and they utilized UCPD in order to define the sequence to construct software requirements.

The practitioners were 25 professionals with at least 2 years of experience who participated in a controlled experiment in 2007 and the quantitative results are detailed in [19]. This experiment was replicated in 2009 with 17 practitioners (with at least 2 years of experience too) who were graduate students of the Master in Informatics program at PUCP. 42 questionnaires filled by practitioners were processed: 25 obtained in the first experiment, and 17 in the second experiment.

2.2 Materials

It was designed a questionnaire which included one question for each constructor of the MAM. Each answer had to be quantified on a five point Likert-type scale [12]. It could be considered as a disadvantage to use only one question for each constructor, but there are some studies that have applied this same approach in other fields such as the medicine with appropriate results [6,8,21]. We wanted to create a user-friendly questionnaire.

The undergraduate students filled the questionnaire at the end of the semester, when they had finished their software projects. The practitioners filled the questionnaire at the end of the controlled experiment in which they were involved into. We commented to the participants that the purpose of the questionnaire is to know their honest opinion about UCPD.

Further details of the questionnaire used and the instruments utilized in the controlled experiment with practitioners can be found at:

http://macareo.pucp.edu.pe/japowsang/precedence/usecase.html

3 Results

The statistical hypotheses to test the working hypothesis H1, H2, and H3 are the following:

H_o: $\mu \leq 3, \alpha = 0.05$

H_a: $\mu > 3$

"μ" is the mean response obtained in the questions related to user's perception about UCPD. We can consider a positive perception of the participants, if the mean response is greater than 3, because a five point Likert-type scale was used in the questionnaries from 1 to 5.

To evaluate the MAM relationships, correlation coeficients and regression analysis were used to formally test hypotheses H4 and H5.

3.1 Perceived Ease of Use

Table 1 presents the results obtained with the question relate to perceived ease of use. It was established a significance level of 0.05 to statistically test the obtained results with undergraduate students and practitioners.

Table 1. Descriptive statistics for perceived ease of use

Variable	Undergraduate students	Practitioners
Observations	31	42
Minimum	3	2
Maximum	5	5
Mean	3.968	3.5
Std. Deviation	0.752	0.891

In order to determine if the obtained results followed a normal distribution, the Shapiro-Wilk test was applied. Since the computed p-values were lower than the significance level α=0.05, the normal distribution hypothesis was rejected for both samples (undergraduate students and practitioners). Due to these results, a parametric test, such as Student's t-test, cannot be used. The Wilcoxon signed rank test was chosen to test the statistical hypothesis defined previously (Ho: $\mu \leq 3$, Ha: $\mu > 3$).

Table 2. Wilkoxon signed rank test results for perceived ease of use

Variable	Undergraduate students	Practitioners
W	253	334
p-value	<0.001	<0.001

Since the computed p-values were lower than the significance level α =0.05, the null hypothesis Ho had to be rejected. It means that we can empirically corroborate working hypothesis H1: the undergraduate students and the practitioners perceived UCPD as easy to use.

3.2 Perceived Usefulness

Table 3 presents the results obtained with the question related to to usefulness. It was established a significance level of 0.05 to statistically test the obtained results.

Table 3. Descriptive statistics for perceived usefulness

Variable	Undergraduate students	Practitioners
Observations	31	42
Minimum	3	2
Maximum	5	5
Mean	4.419	3.905
Std. Deviation	0.7199	0.878

In order to determine if the obtained results followed a normal distribution, we applied the Shapiro-Wilk test. Since the computed p-values were lower than the significance level α=0.05, the normal distribution hypothesis was rejected for both samples (undergraduate students and practitioners). Due to these results, a parametric test, such as Student's t-test, cannot be used. The Wilkoxon signed rank test was chosen to test the statistical hypothesis defined previously (Ho: $\mu \leq 3$, Ha: $\mu > 3$).

Table 4. Wilkoxon signed rank test results for perceived usefulness

Variable	Undergraduate students	Practitioners
W	378	526.5
p-value	<0.001	<0.001

Since the computed p-values were lower than the significance level α =0.05, the null hypothesis Ho had to be rejected. It means that we can empirically corroborate working hypothesis H2: the undergraduate students and the practitioners perceived UCPD as useful.

3.3 Intention to Use

Table 5 presents the results obtained with the question relate to intention to use. It was established a significance level of 0.05 to statistically test the obtained results.

Table 5. Descriptive statistics for perceived usefulness

Variable	Undergraduate students	Practitioners
Observations	31	42
Minimum	2	1
Maximum	5	5
Mean	4.226	3.738
Std. Deviation	0.805	1.014

In order to determine if the obtained results followed a normal distribution, we applied the Shapiro-Wilk test. Since the computed p-values were lower than the significance level α=0.05, the normal distribution hypothesis was rejected for both samples (undergraduate students and practitioners). Due to these results, the Wilkoxon signed rank test was chosen to test the statistical hypothesis defined previously (Ho: $\mu \leq 3$, Ha: $\mu > 3$).

Table 6. Wilkoxon signed rank test results for perceived usefulness

Variable	Undergraduate students	Practitioners
W	370.5	457
p-value	<0.001	<0.001

Since the computed p-values were lower than the significance level α =0.05, the null hypothesis Ho had to be rejected. It means that we can empirically corroborate working hypothesis H3: the undergraduate students and the practitioners have the intention to use UCPD.

3.4 MAM Evaluation

To assess the relationships between variables proposed in the MAM, we must use the correlation coefficient, similar to the studies conducted by Davis [9] and Adams et. al [2].

According Muijs [15] in order to determine if there is a degree of relationship between two ordinal variables, the Spearman's correlation coefficient must be used (not the Pearson's correlation one). The Likert-scale used in the questionnaires is ordinal, for this reason Spearman's correlation (Spearman's rho) had to be used to evaluate MAM. The rules of thumb to determine the strength of a relationship proposed by Muijs are the following:

<0. +/-1 weak
<0. +/-3 modest
<0. +/-5 moderate
<0. +/-8 strong
>= +/-0.8 very strong.

Fig.2 presents Spearman's rho (ρ) and the strength for each relationship. It can be observed that usefulness and ease of use were significantly correlated each other for practitioners(ρ=0.549, p-value=0.000) but they were not correlated each other for undergraduate students. Usefulness and intention to use were significantly correlated each other for both type of participants (students and practitioners). Ease of use and intention to use were only significantly correlated each other for practitioners. It means that we can empirically corroborate working hypotheses H4 and H5 using the practitioners sample. Unfortunately, working hypotheses H4 and H5 can not be corroborated for the undergraduate students sample.

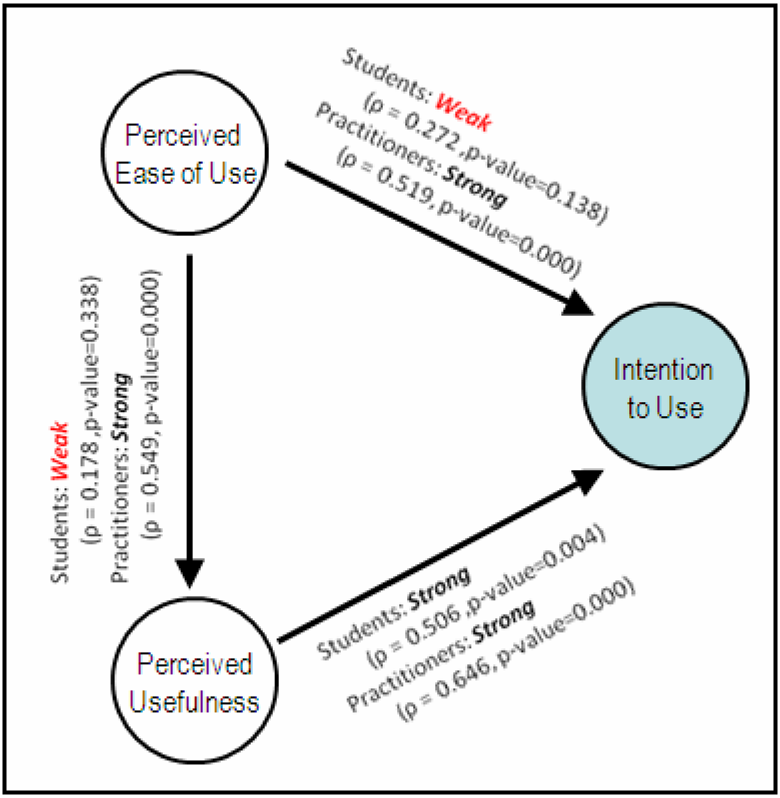

Perceived Ease of Use
Intention to Use
Perceived Usefulness
Students: Weak (ρ = 0.272, p-value=0.138)
Practitioners: Strong (ρ = 0.519, p-value=0.000)
Students: Weak (ρ = 0.178, p-value=0.338)
Practitioners: Strong (ρ = 0.549, p-value=0.000)
Students: Strong (ρ = 0.506, p-value=0.004)
Practitioners: Strong (ρ = 0.646, p-value=0.000)

Fig. 2. Spearman's rho and strength for MAM's relationships

In order to confirm the causal relationships between variables defined in the MAM, a regression analysis had to be used , similar to the study included in Abrahao's Phd thesis [1], for those variables with a correlation coefficient equal or better than "moderate". Because the results obtained are ordinal, an ordinal regression model had to be used. This kind of regression model was proposed by McCullagh and Nelder [12].

Ordinal regression model for perceived usefulness and perceived ease of use. In order to determine the ordinal regression model for the practitioners sample, it was used the Perceived Ease of Use (PEU) as the independent variable and the Perceived Usefulness (PU) as the dependent variable. Table 7 shows obtained ordinal regression model for PEU and PU, calculated using SPSS software, using the *Probit* link function.

Using Table 7 and the definitions made by Borooah [4], if we want to obtain the perceived usefulness based on the perceived easy of use, we need to calculate the following: ***D = -0.796 x PEU.*** The PU was determined with the D calculated:

- PU = 2, if D $\leq$ -1.017
- PU = 3, if -1.017 $\leq$ D $\leq$ -2.072
- PU = 4, if -2.072 $\leq$ D $\leq$ -3.583
- PU = 5, if D $\geq$ -3.583

Table 7. Parameter estimates for Perceived Usefulness - Perceived Ease of Use

		Estimate	Std. Error	df	p-value
Threshold	[PU = 2.00]	1.017	.733	1	.165
	[PU = 3.00]	2.072	.755	1	.006
	[PU = 4.00]	3.583	.850	1	.000
Location	PEU	.796	.219	1	.000

The model chi-square is 14.101 with 2 degrees of freedom and p-value =0.000. This is highly significant, and the model confirms that perceived ease of use has a significant effect on the perceived usefulness for the practitioners.

Table 8. Test of parallel lines Perceived Usefulness - Perceived Ease of Use

Model	-2 Log likelihood	Chi-square	df	p-value
Null hypothesis	27.361			
General	27.143	0.218	2	0.897

For location-only models, the test of parallel lines assesses if the assumption that the parameters are the same for all categories is reasonable. Table 8 presents the test of parallel lines for the calculated model. As it can be observed in the above table, the chi-square value is insignificant and the p-value is greater than 0.05. It means the ordinal regression model calculated accomplish with the required assumption.

In the linear regression model, the coefficient of determination, R^2, summarizes the proportion of variance in the dependent variable associated with the predictor variables (independent variables). For ordinal regression models, it is not possible to compute a single R^2 statistic, so these approximations of a R^2 are computed instead.

Table 9. R^2 calculated for Perceived Usefulness - Perceived Ease of Use

Type of R^2	Value
Cox and Snell	0.285
Nagelkerke	0.312
McFadden	0.137

The results presented in Table 9 indicate that the calculated ordinal regression model explains between 31.2% and 13.7% of the variability of the perceived usefulness. It is important to note that these values must be interpreted with caution, since they are not direct equivalents to the R^2 statistics obtained in a linear regression model.

The regression model presented confirms the working hypothesis H4. This means the perceived usefulness is determined by the perceived ease of use for the practitioners sample.

Ordinal regression model for intention to use vs. perceived usefulness and perceived ease of use. In order to determine the ordinal regression model for the practitioners sample, the Perceived Ease of Use (PEU) and the Perceived Usefulness (PU) were used as dependent variables and the Intention to Use (IU) as the independent (predictor) variable. Table 7 shows obtained ordinal regression model for IU vs PEU and PU, calculated in SPSS software, using the *Probit* link function.

Table 10. Parameter estimates for Intention to Use vs. Perceived Usefulness and Ease of Use.

		Estimate	Std. Error	df	p-value
Threshold	[IU = 1.00]	1.868	.937	1	.046
	[IU = 2.00]	2.762	.925	1	.003
	[IU = 3.00]	3.898	.986	1	.000
	[IU = 4.00]	5.472	1.112	1	.000
Location	PEU	.415	.234	1	.076
	PU	.764	.248	1	.002

Using Table 10 and the definitions made by Borooah [4], if we want to obtained the perceived usefulness based on the perceived easy of use, we need to calculate the following: ***D = -0.415 x PEU + 0.764 x PU.*** The IU was determined with the D calculated:

- IU = 1, if D $\leq$ 1.868
- IU = 2, if 1.868$\leq$ D $\leq$ 2.762
- PU = 3, if 2.762$\leq$ D $\leq$ 3.898
- PU = 4, if 3.898 $\leq$ D $\leq$ 5.472
- PU = 5, if D $\geq$ 5.472

The model chi-square is 21.906 with 2 degrees of freedom and p-value =0.000. This is highly significant, and the model confirms that perceived ease of use and perceived usefulness have a significant effect on the intention to use UCPD for the practitioners.

Table 11. Test of parallel lines: Intention to Use vs. Perceived Usefulness and Ease of Use

Model	-2 Log likelihood	Chi-square	df	p-value
Null hypothesis	53,521			
General	46,951	6,570	2	0.362

Table 11 presents the test of parallel lines for the calculated model. As it can be observed in the above table, the chi-square value is low and the p-value is greater than 0.05. It means the ordinal regression model calculated accomplish with the required assumption.

In the linear regression model, the coefficient of determination, R^2, summarizes the proportion of variance in the dependent variable associated with the predictor variables (independent variables). For ordinal regression models, it is not possible to compute a single R^2 statistic, so these approximations of a R^2 are computed instead.

Table 12. R^2 calculated for Intention to Use vs. Perceived Usefulness and Perceived Ease of Use

Type of R^2	Value
Cox and Snell	0,406
Nagelkerke	0,435
McFadden	0,191

The results presented in Table 12 indicate that the calculated ordinal regression model explains between 40.6% and 19.1% of the variability of the perceived usefulness. It is important to note that these values must be interpreted with caution, since they are not direct equivalents to the R^2 statistics obtained in a linear regression model.

The regression model presented confirms the working hypothesis H5. This means the intention to use is determined by the perceived ease of use and the perceived usefulness for the practitioners sample.

4 Conclusions and Future Work

This paper describes an empirical study that evaluates the intention to use the UCPD technique that is used to determine software construction sequences taking into account the developer's perspective. The study considers the perceptions of undergraduate students and practitioners with at least 2 years of experience in software development projects.

UCPD is perceived as easy to use and useful for all of the participants (undergraduate students and practitioners). Also, the participants of this study acknowledged having the intention to use UCPD in next software development projects. These results do not disagree with the quantitative results obtained in the controlled experiment with practitioners (previously published [19]) and the replicated experiment.

Although the perceptions of UCPD are positive for all of the participants, the relationships defined in the MAM are only confirmed with the statistical tests applied using the practitioners' sample.

Many researchers comment the benefits to use undergraduate students for research studies. However, it should be noted that in some situations, similar to this study, the results obtained with undergraduate students should be taken with caution and it is preferable to use practitioners with experience in the industry, in order to get confident results.

As a future work, we plan to replicate the controlled experiment with undergraduate students in order to contrast and confirm the results obtained in this study.

Acknowledgments

This research work has been performed with the support of Dirección Académica de Investigación of Pontificia Universidad Católica del Perú, under projects DAI-E039 and DAI-4051.

References

1. Abrahão, S.: On the Functional Size Measurement of Object-Oriented Conceptual Schemas: Design and Evaluation Issues, PhD Thesis, Department of Information Systems and Computation, Valencia University of Technology (October 2004)
2. Adams, D., Nelson, R., Todd, P.: Perceived usefulness, ease of use, and usage of information technology: a replication, MIS Quarterly, USA (1993)
3. Basili, V.R., Caldiera, G., Rombach, H.D.: Goal Question Metric Paradigm. In: Marciniak, J.J. (ed.) Encyclopedia of Software Engineering. Wiley, Chichester (1994)
4. Borooah, V.K.: Logit and Probit: Ordered and Multinomial Models. Sage Publications, USA (2001)
5. Carver, J., Jaccheri, L., Morasca, S.: Issues in Using Students in Empirical Studies in Software Engineering Education. In: METRICS 2003, p. 239. IEEE Computer Society, USA (2003)
6. Cepeda, M.S., Chapman, C.R., Miranda, N., Sanchez, R., Rodriguez, C.H., Restrepo, A.E., Ferrer, L.M., Linares, C.D.B.: Emotional Disclosure Through Patient Narrative May Improve Pain and Well-Being: Results of a Randomized Controlled Trial in Patients with Cancer Pain. Journal of Pain and Symptom Management 35(6), 623–631 (2008)
7. Condori, N.: Un Procedimiento de Medición de Tamaño Funcional para Especificaciones de Requisitos, PhD Thesis, Department of Information Systems and Computation, Valencia University of Technology (2007)
8. Davey, H.M., Barratt, A.L., Butow, P.N., Deeks, J.J.: A one-item question with a Likert or Visual Analog Scale adequately measured current anxiety. Journal of Clinical Epidemiology 60, 356–360 (2007)
9. Davis, F.D.: Perceived Usefulness, Perceived Ease of Use and User Acceptance of Information Technology. MIS Quarterly, 319–340 (1989)
10. IBM Corporation, Rational Unified Process version 2001A.04.00.13, USA (2001)
11. Jacobson, I.: Object-Oriented Software Engineering. A Use Case Driven Approach. Addison-Wesley, USA (1992)
12. Likert, R.: A technique for the measurement of attitudes. Archives of Psychology. Columbia University Press, New York (1931)
13. McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edn. Chapman & Hall, London (1989)
14. Moody, D.L.: Dealing with Complexity: A Practical Method for Representing Large Entity Relationship Models, PhD. Thesis, Department of Information Systems,University of Melbourne, Australia (2001)
15. Muijs, D.: Doing Quantitative Research in Education with SPSS. Sage Publications, USA (2004)
16. Object Management Group, OMG Unified Modeling Language, USA (2008), http://www.uml.org

17. Poels, G., Maes, A., Gailly, F., Paemeleire, R.: Measuring User Beliefs and Attitudes towards Conceptual Schemas: Tentative Factor and Structural Equation Model. In: Fourth Annual Workshop on HCI Research in MIS (December 2005)
18. Pow-Sang, J.A., Nakasone, A., Imbert, R., Moreno, A.M.: An Approach to Determine Software Requirement Construction Sequences based on Use Cases. In: Proceedings Advanced Software Engineering and Its Applications ASEA 2008, Sanya, China. IEEE Computer Society, Los Alamitos (2008)
19. Pow-Sang, J.A., Nakasone, A., Moreno, A.M., Imbert, R.: Evaluating the Applicability of a Use Case Precedence Diagram based Approach in Software Development Projects through a Controlled Experiment, Advances in Security Technology (Revised selected papers of SecTech 2008, Communications in Computer and Information Science (CCIS), LNCS, Springer,Heidelberg)
20. Rosenberg, D., Scott, K.: Use Case Driven Object Modeling with UML. Addison-Wesley, Massachusets (1999)
21. Temel, J.S., Pirl, W.F., Recklitis, C.J.: Feasibility and validity of a one-item fatigue screen in a thoracic oncology clinic. Journal of Thoracyc Oncology 1(5) (June 2006) Lippincott Williams & Wilkins

A Simple Method Using Multi-Core and Multiple GbE Ports for Improving Parallel Performance in Commodity PC Cluster

Kei Shimada, Miku Kai, Tsuyoshi Yoshioka, and Takafumi Fukunaga

Kumamoto Prefectural College of Technology, 4455-1 Kikuyo-cho,
Kikuchi-gun, Kumamoto, 869-1102 Japan
t-fukunaga@kumamoto-pct.ac.jp

Abstract. Due to advent of powerful and easily available Multi-Core PC clusters, the computation performance of each node is dramatically increased and this trend will probably continue in the future. On the other hand, the use of powerful network systems (Myrinet, Infiniband, etc.) is expensive and tends to increase the difficulty of programming and degrades portability because they need dedicated libraries and protocol stacks. This paper proposes portable method to improve bandwidth-oriented parallel applications by improving the bandwidth performance without the above dedicated hardware, libraries, protocol stacks and IEEE802.3ad (LACP). Since the proposed method is introduced only by loading the proposed driver without modifying TCP/IP protocol stacks and existing applications, it has advantages in both high portability and stability. The proposed method also performs better than IEEE802.3ad without LACP switches and LACP drivers. Moreover LACP performance is influenced by the network parameters (MAC addresses, IP addresses, VLAN id, etc.) because its distribution algorithm uses these parameters. On the other hand, the proposed method shows the stable effect regardless of these parameters.

Keywords: Bandwidth, Parallel processing, Bonding driver, Multi-Core, Multi GbE port.

1 Introduction

In recent years, since Multi-Core PCs have become more common and inexpensive, the use of PC clusters employing them is now widespread among the business and research world and this trend will probably continue in the future. The computing performance increases in direct relation with the number of cores of each node. On the other hand, Gigabit Ethernet (GbE, hereafter) has been used for interconnection networks of commodity PC clusters more than ten years because of its high cost-effectiveness, that is, during that time the communication performance is almost the same. This disproportion in performance between nodes and networks in commodity PC clusters easily causes communication bottleneck for the parallel applications that transfer a large amount of data. Most of the studies that improve communication performance in PC

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 538–548, 2010.

clusters are adopting dedicated hardware and its dedicated protocol (Myrinet [1]-[2], Infiniband [3]-[5], Quadrics Network [6], DIMMnet-2 [7], RHiNET [8]-[10], PACS-CS [11]), non standard protocol (PM [12], BIP [13], FM [14], VMMC[15]) and dedicated library [16]. The main issue of them is the increase in cost that includes not only hardware price but also porting work cost.

In this paper, the method MCMGP (Multi-Core using Multi-Gigabit-Port) is proposed to improve the performance of parallel applications which need high bandwidth without the above dedicated hardware, libraries and protocol stacks for clusters. The proposed method is introduced to Multi-Core PC clusters by using only multiple GbE ports and the driver implementing the function as described in later Section 3. Since this driver is a loadable module, it is easy to introduce and easy to remove. The main advantage of MCMGP is high portability because what it needs are just Multi-Core PC nodes, Multi Gigabit Ethernet ports and the loadable driver proposed in this paper. Both devices above have already become commodity ones. The introduction of MCMGP does not require any modification in applications, operating system and NIC drivers. When using Multi-Core PC cluster with Multi Gigabit ports, the needed operation to introduce MCMGP is simply to load the driver. For some reason when the original environment without MCMGP is requested, the needed operation is simply to unload the driver and to restart network interface configuration. In addition, the driver is inserted between existing protocol stack and existing NIC drivers, and executing program code on transmitting frames is only about 50 lines, those on receiving frames is zero. This means all applications are executed by almost the same protocol stacks code and NIC driver code as before. Accordingly, the second advantage of MCMGP is to keep almost the same stability as before. On the other hand, there is no denying that the dedicated systems using relatively new protocol stacks and libraries for dedicated hardware are subject to instability compared to the usual systems using TCP/IP improved through many years and frequently used libraries.

There is IEEE802.3ad (LACP) which is similar to MCMGP in respect to using multiple Ethernet ports. This protocol makes a logical communication port from multiple physical ports both for improving the bandwidth performance and changing to spare circuit when a circuit fault occurs. When LACP is used for improving the bandwidth, it generally improves the bandwidth between the server and the switch and between the switches. Accordingly, LACP improves the one-to-many communication performance of clients-server systems, but does not improve the one-to-one communication performance as described in subsection 4.1. In addition, the LACP supported switches and supported NIC drivers are mandatory. On the other hand, MCMGP works on any switch and any NIC driver. This advantage indicates again that MCMGP has high portability and high cost-effectiveness. The experimental results of the bandwidth and parallel application performance show MCMGP achieves better performance than LACP.

The contents of this paper are organized as follows: In Section 2, the design of MCMGP is described. Section 3 sketches the implementation. Experimental results are presented in Section 4. The paper ends with a conclusion and references.

2 Proposed System

The aim of this proposal is to improve the performance of high bandwidth parallel applications by the use of multiple GbE ports and a loadable driver. Fig.1 illustrates how the sender allocates a dedicated GbE port to each of high bandwidth streams on the nodes using 4 cores (processors) PC equipped with 4 GbE ports. Although a proposed method needs the ports of the same or large number as cores, procuring them is getting easier with the advent of inexpensive Multiple ports network interfaces.

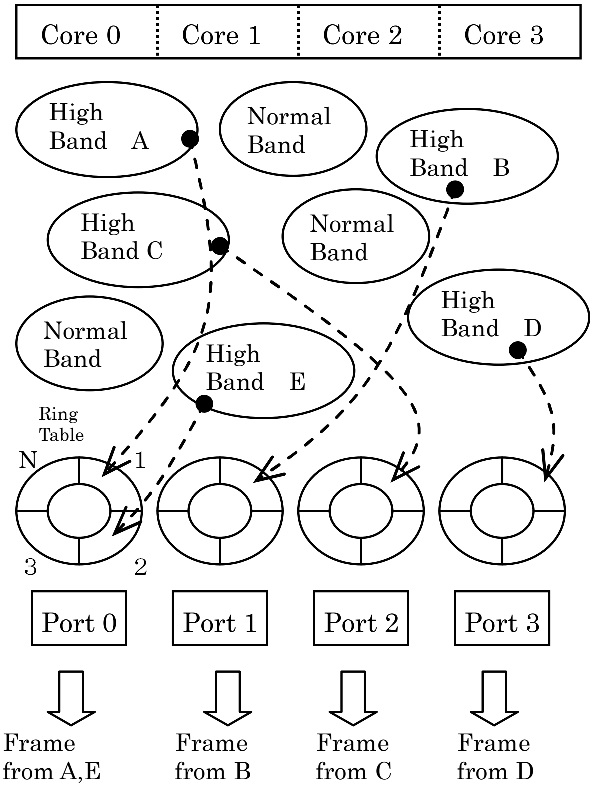


Fig. 1. Outline of proposed method on the nodes using 4 cores PCs equipped with 4 GbE ports. High Band stands for the high bandwidth stream. Port stands for GbE port. N denotes the number of entries of each Ring Table.

As shown in Fig.1, each of the ports is associated with the newly appended Ring Tables on a one-to-one correspondence basis. Each table keeps TCP port numbers of the flames to send from the associated GbE port. A proposed driver registers TCP port numbers of High bandwidth streams to these tables on a round-robin basis as shown by High Band A to E in Fig.1. The decision as to whether a stream is high bandwidth or not is determined by the amount of send data left on the send buffer which didn't complete transfer. Complete transfer means receiving acknowledgement for sent data from the receiver. The amount of send data left is calculated by the subtraction of the first byte we want to acknowledge from the last byte of data to send which has been transferred from the user applications. If the amount of data left on the send buffer is larger than the threshold, its stream is regarded as high bandwidth. This condition is based on the expectation that the high bandwidth TCP streams leave lots of data to send on the send buffer. The data left are composed of those that are waiting for acknowledgement from the receiver and those that have not been sent outside yet after being transferred from the application. With a threshold value of 7000, streams greater than 250Mbits/s approximately are regarded as high bandwidth. So, the value of the threshold is implemented at 7000. After registering TCP port number as entry of the Ring Table, its stream is sent from the associated GbE port until the entry of the Ring Table, if the Ring Table sequence re-starts the data is overwritten. Naturally the overwritten entry is registered again as long as its bandwidth continues to be high bandwidth.

As shown in Fig.1, N denotes the number of entries of each Ring Table. The implemented value of N is 3 because of the following reason. The number of high bandwidth streams which one port can send simultaneously is 3 or 4 at the most because of hardware performance limitations (1Gbits/s), since the registered stream's speed is more than 250Mbits/s. The value of N is implemented at 3 for that reason. Even if there are more streams to register than expected, the proposed method can easily increase the number of ports by installing the Multiple port network interfaces.

Fukunaga et al. [17] modified GbE destination MAC addresses to divide the communication paths completely as shown in Fig.2, and is also implemented in this proposal. Each of the ports at the sender node is associated with the ports at the receiver node on a one-to-one correspondence basis with a MAC table (see Section 3, Table 1) prepared in advance. According to its correspondence, destination MAC addresses are replaced just before sending. As a result of address replacement, the frames from port *N* of the sender reach to port *N* of the receiver.

The proposed method has been achieved by modifying an existing Bonding driver code at the sender PC. There are no added coding at the receiver. Although the one overhead presented is look up time on the Ring Table at the sender, even the N value of 10 only degrade performance of the total bandwidth by 1~2 percent. This show how small the overhead is.

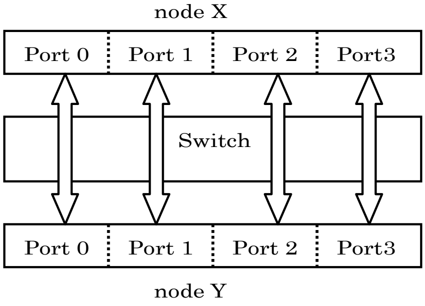


Fig. 2. Illustration of the transfer with one-to-one correspondence between the sender's ports and the receiver's ports by a method in previous paper [17]

3 Implementation

The proposal is implemented by modifying the existing Bonding driver in the Linux operating system that provides fault tolerance and communication balancing functions by using multiple Ethernet ports. This existing driver has seven modes including round-robin load balancing and IEEE802.3ad (LACP). A proposed method has been implemented in the program codes of round-robin function in this driver.

Fig.3 shows how the high bandwidth streams are registered to Ring Tables. TCP/IP frames are identified by the protocol field of the IP header ((1) in Fig.3). Non-TCP/IP frames are processed by original program code of round-robin function (2). If one of

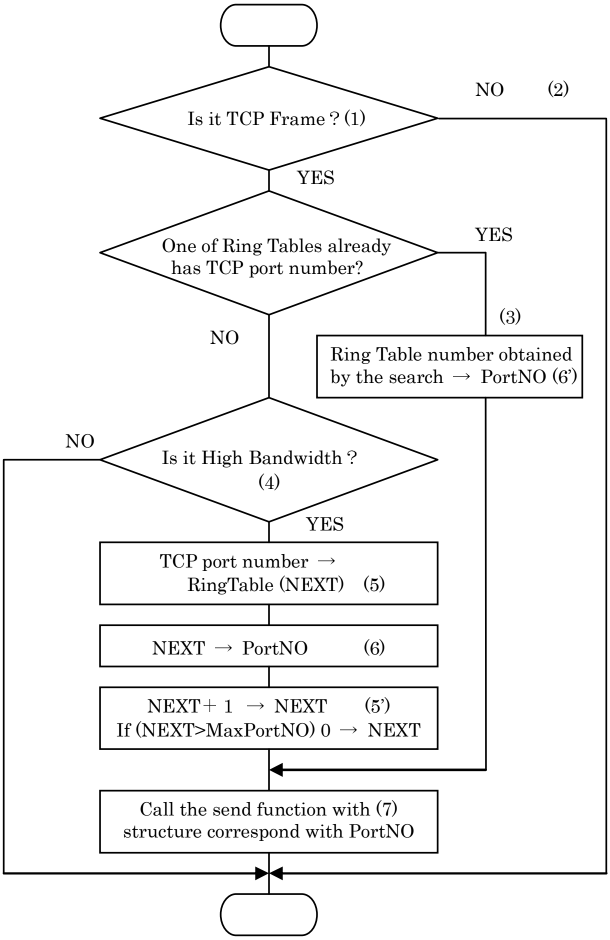

Is it TCP Frame ? (1)
NO
(2)
YES
One of Ring Tables already has TCP port number?
YES
(3)
NO
Ring Table number obtained by the search → PortNO (6')
NO
Is it High Bandwidth ? (4)
YES
TCP port number → RingTable (NEXT) (5)
NEXT → PortNO (6)
NEXT+ 1 → NEXT (5')
If (NEXT>MaxPortNO) 0 → NEXT
Call the send function with (7) structure correspond with PortNO

Fig. 3. Flow chart of algorithm which only registers TCP port number of high bandwidth streams. Ring Tables are indexed by *NEXT*. *PortNO* get selected number of GbE ports to send the frame. *MaxPortNO* denotes max GbE port number.

the Ring Tables already has TCP port number of the frame, GbE port which is associated with the table is selected as the send port (3) ; if not, the frame proceeds to the next block to make a decision as to whether to be high bandwidth stream or not by the condition mentioned in previous section (4). The TCP port number, which is regard as high bandwidth, is registered to the next Ring Table on a round-robin basis (5), (5'). In the end this algorithm assigns selected GbE port number to PortNO-variable (6), (6'). Then the frame is sent using the corresponding net_device structure of PortNO-variable which has information about the selected port (7).

In addition, just before sending a frame, both Ethernet addresses (i.e. destination and source MAC address) are modified to transfer with one-to-one communication form as shown in Fig.2 by a method described in [17]. The frames sent from the port *N* of the sender always reach the port *N* of the receiver. The program code for the addresses modification is also implemented in Bonding driver. The association between the sender port and the receiver port is described in MAC-table as shown in Table 1. This table, which has Ethernet MAC address information of all of cluster nodes, has to be prepared in advance by using registration tool. *Destination node key* in Table 1 is MAC address of the first port of each node and search-key to identify the receiver. Since the MAC address of the original send frame is one of the first port of the receiver, this table is looked up by its address to find the entry of the receiver. *Port number key* is the consecutive number of the ports of each node and search key to identify the receive port. Naturally the receive port number correspond to the send port number. Original destination MAC address is replaced by *New destination MAC* obtained from looking up the entry by these keys. Since source MAC address of the original frames sent from any port is the same as the first port when the existing Bonding driver is used, the modified code also replace it by the actual MAC addresses of send port.

Table 1. Outline of MAC-table

	Explanation
Destination node key (index 1)	Destination PC MAC-address
Port number key (index 2)	Port consecutive numbers
New destination MAC	MAC address of the receiver port corresponding with the sender port

4 Evaluation

This section evaluates the communication capabilities (bandwidth and latency) and the parallel processing performance of the proposed method against IEEE802.3ad, round-robin load balancing, and normal transmission (i.e. one GbE port).

For each benchmark, following 4 methods have been tried: 1) a normal transmission using one GbE port (labeled as 1 Port), 2) round-robin load balancing using round-robin mode of the existing Bonding driver (labeled as Round-robin), 3) IEEE802.3ad (LACP) using IEEE802.3ad mode of the existing Bonding driver (labeled as LACP), 4) proposed method (labeled as Proposal). The hosts and switch in the testbed are Linux 2.6.24 systems running on Double Dual-Core 2.4GHz PCs (4 cores per node) each with 6GB of SDRAM and with 5 Gigabit Ethernet Interfaces (Intel PRO/1000, nVidia MCP55) and NETGEAR GSM7248R Gigabit switch. The testbed is a cluster of 9 of these Multi-Core nodes in which each node has 4 cores and 5 Gigabit ports connected by 5 UTP cables to the Gigabit switch.

Since implementations of distribution algorithm of LACP vary in parameters (e.g. MAC address, VLAN id, IP address, etc.), the load balancing condition depends on the environments. This time the switch adopts TCP port number based distribution algorithm from the 6 types it has, and NIC drivers (existing Bonding drivers) adopt MAC address based distribution algorithm which is an only implementation.

4.1 Bandwidth and Latency

The bandwidth and the latency are evaluated for above 4 methods with Netperf-1.2.7 benchmark. Fig.4 and Fig.5 shows a total of bandwidth between two nodes using 4 ports and 5 ports respectively, each of them executes 4 benchmarks in order to measure under the same conditions as parallel processing using 4 cores per node. On the total, Proposal achieves the best bandwidth for almost any message sizes except small messages (equal to or smaller than 16 Bytes). Under this experimental environment for measuring bandwidth, LACP can not obtain the speedup.

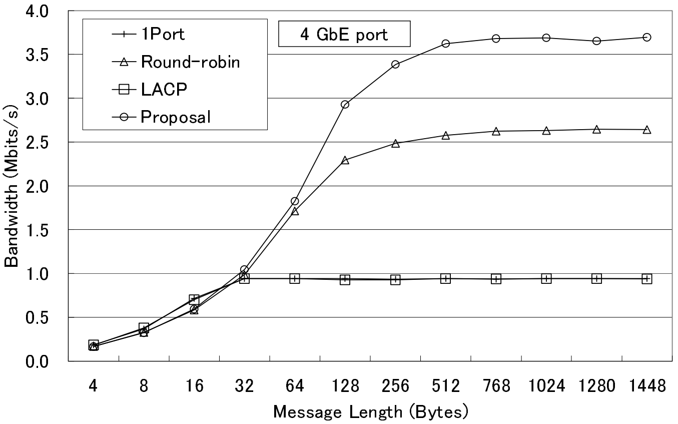


Fig. 4. Bandwidth of data transfers between nodes using 4 GbE ports. *1 NIC* stands for normal transfer with 1 NIC, *Round-robin* stands for round-robin load balancing with existing Bonding driver's round-robin mode, *LACP* stands for IEEE802.3ad with existing Bonding driver's LACP mode, *Proposal* stands for proposed method.

Round-robin also is far lower than Proposal because of a large quantity of SACK frames. SACK frames are sent to the sender by the receiver to inform out-of-order of the frame sequence. Due to SACK frames the sender has received one and a half times more acknowledgement frames including SACK than Proposal in spite of lower

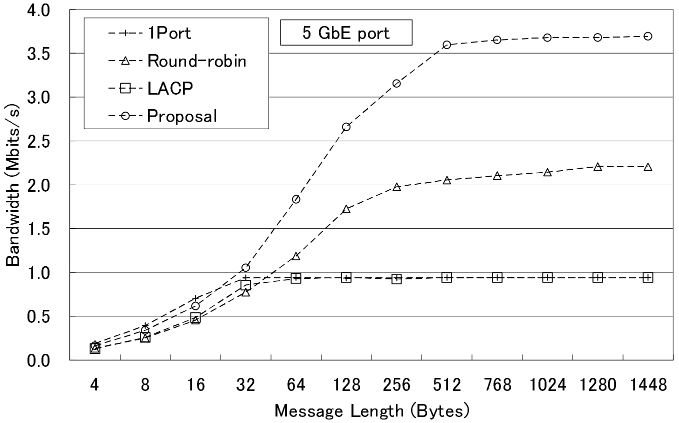


Fig. 5. Bandwidth of data transfers between nodes using 5 GbE ports. *1 NIC* stands for normal transfer with 1 NIC, *Round-robin* stands for round-robin load balancing with existing Bonding driver's round-robin mode, *LACP* stands for IEEE802.3ad with existing Bonding driver's LACP mode, *Proposal* stands for proposed method.

performance. In proposal method, since each of high bandwidth streams has the only communication path, the rate of SACK frames is about only 1 percent of the total acknowledgement frames. A slightly drop in small size are due to overhead of additional code, but this is not an important matter as my targets are high bandwidth streams.

Considering the increase of the number of GbE ports, although Round-robin degrades the speedup by 17 percent due to the increase of overhead, Proposal shows the same speedup as shown in Fig.5. This mean a proposal is easy to increase the ports in terms of overhead.

On the other hand, the results of latency measurement are 53.0 (1 Port), 55.5 (Round-robin), 53.7 (LACP), 61.3 (Proposal) micro-seconds. The results show a drop in latency due to the overhead. Although this gives apprehension in the applications which dominant factor is latency performance, the proposed method gives higher priority to stability of TCP/IP than adopting dedicated low latency protocols like PM, FM, VMMC etc.

4.2 Parallel Processing Performance

The parallel processing performances of above 4 methods are evaluated with FT, MG, IS and CG class B in NAS Parallel Processing Benchmark (NPB) 3.3 which send a large amount of data per second. Fig.6, 7, 8 and 9 show the results. External communication is occurred with more than 8 processors. Proposal can scale up very well, obtaining better speedups than the others. Compared to LACP, Proposal achieves a 29 percent improvement in performance for FT and a 10 percent improvement for MG with 32 processors (cores).

On the contrary, all methods can not scale up at all in IS and CG because the dominant factor in them is the latency performance. The case of 4 processors shows the best performance on the total since it doesn't need external transfers.

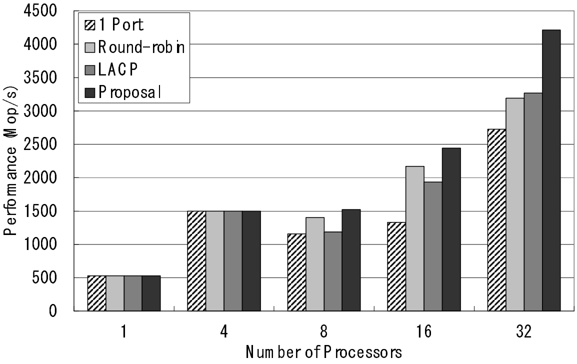


Fig. 6. FT class B benchmark in NAS Parallel Benchmarks (NPB)

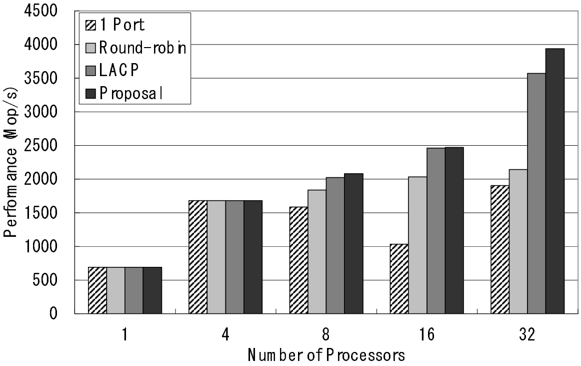


Fig. 7. MG class B benchmark

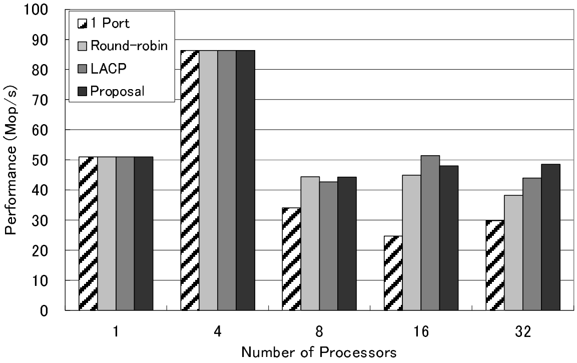


Fig. 8. IS class B benchmark

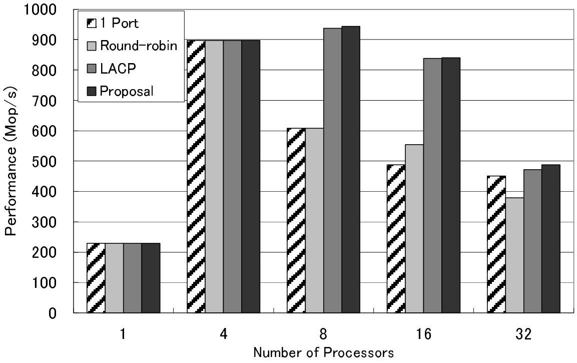


Fig. 9. CG class B benchmark

5 Conclusion

This paper propose a method for the improvement of the bandwidth without using expensive hardware (Myrinet, Infiniband etc.), dedicated libraries and non standard protocols tending to affect cost and complexity of programming. It is obvious from the experimental results that this proposal has a good influence on the bandwidth-oriented parallel applications.

IEEE802.3ad is similar to the proposal in respect to distributing data frames to multiple ports to obtain the improvement in communication capability. However the proposal allocates each of high bandwidth streams to a fixed path to avoid out-of-order frame sequences without being influenced by the environment. Moreover the proposal is more flexible because it is achieved by only the sender driver's control.

References

1. Boden, N.J., Cohen, D., Felderman, R.E., Kulawik, A.E., Seitz, C.L., Seizovic, J.N., Su, W.-K.: Myrinet: A gigabit-per-second local area network. IEEE Micro., 29–36 (February 1995)
2. Tezuka, H., O'Carroll, F., Hori, A., Ishikawa, Y.: Pin-down Cache: A Virtual Memory Management Technique for Zero-copy Communication. In: 12th IPPS and 9th SPDP, Orlando, FL (March 1998)
3. InfiniBandTM Architecture Specification, InfiniBand Trade Association (2004), `http://www.infinibandta.org`
4. Gangadharappa, T., Koop, M., Panda, D.K.: Designing and Evaluating MPI-2 Dynamic Process Management Support for InfiniBand. In: International Conference on Parallel Processing Workshops, pp. 89–96 (September 2009)
5. Lin, Y., Han, J., Gao, J., He, X.: uStream: A User-Level Stream Protocol over Infiniband. In: 15th International Conference on Parallel and Distributed Systems, pp. 65–71 (December 2009)
6. Petrini, F., Fang, W.-C., Hoisie, A., Coll, S., Frachtenberg, E.: The Quadrics Network: High-Performance Clustering Technology. IEEE Micro 22(1), 46–57 (2002)

7. Kitamura, A., Hamada, Y., Miyabe, Y., Izawa, T., Miyasiro, T., Tanabe, N., Nakajo, H., Amano, H.: Design and Implementation of Network Interface Controller on DIMMnet-2. Trans. IPSJ 46(SIG 12), 13–26 (2005)
8. Kudoh, T., Yamamoto, J., Nishi, H., Nishimura, S., Tatebe, O., Amano, H.: RHiNET: A Network for High Performance Parallel Computing Using Locally Distributed Computers, pp. 69–73 (November 1999)
9. Nakajo, H., Ishii, M., Yamamoto, J., Kudo, T., Yokoyama, T., Tsuchiya, J., Amano, H.: Cache Coherence Protocol for Home Proxy Cache on RHiNET and its Preliminary Performance Estimation. In: Innovative Architecture for Future Generation High-Performance Processors and Systems (IWIA 2001), pp. 53–60 (January 2001)
10. Watanabe, K., Otsuka, T., Tsuchiya, J., Amano, H., Harada, H., Yamamoto, J., Nishi, H., Kudoh, T.: Performance Evaluation of RHiNET-2/NI: A Network Interface for Distributed Parallel Computing Systems. In: Third IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2003), pp. 318–325 (May 2003)
11. Boku, T., Sato, M., Ukawa, A., Takahashi, D., Sumimoto, S., Kumon, K., Moriyama, T., Shimizu, M.: PACS-CS: A Large-Scale Bandwidth-Aware PC Cluster for Scientific Computations. In: Sixth IEEE International Symposium on Cluster Computing and the Grid (CCGRID 2006), pp. 233–240 (May 2006)
12. Tezuka, H., Hori, A., Ishikawa, Y.: PM: a high-performance communication library for multi-user parallel environments. Technical Report TR-96015, Real World Computing Partnership (1996)
13. Prylli, L., Tourancheau, B.: BIP: a new protocol designed for high performance. In: In PC-NOW Workshop, held in parallel with IPPS/SPDP 1998, Orlando, USA, March 30-April 3 (1998)
14. Pakin, S., Lauria, M., Chien, A.: High performance messaging on workstations: Illinois Fast Messages (FM) for myrinet. In: Supercomputing 1995 (1995)
15. Dubnicki, C., Bilas, A., Chen, Y., Damianakis, S., Li, K.: VMMC-2: efficient support for reliable, connection-oriented communication. In: Proceedings of Hot Interconnects (August 1997)
16. Araki, S., Bilas, A., Dubnicki, C., Edler, J., Konishi, K., Philbin, J.: User-Space Communication: A Quantitative Study. In: Proceeding of the 1998 ACM/IEEE SC 1998 Conference (1998)
17. Fukunaga, T., Umeno, H.: Implementation and evaluation of improvement in parallel processing performance on the cluster using small-scale SMP PCs. Trans. IEE Japan 128(12), 1842–1851 (2008)

Proposal for Sophisticated Periodic Execution Control in Embedded Systems

Yuuki Furukawa, Toshihiro Yamauchi, and Hideo Taniguchi

Graduate School of Natural Science and Technology, Okayama University
furukawa@swlab.cs.okayama-u.ac.jp,
{yamauchi,tani}@cs.okayama-u.ac.jp

Abstract. In embedded systems, the types of processings to be executed are limited, and many processes are executed periodically. In such systems, we need to reduce the overhead of periodic execution control and the dispersion of the processing time. ART-Linux has been proposed as a conventional real-time operating system that can be used for this purpose in various devices such as robots. In this paper, we discuss the periodic execution control of ART-Linux and clarify several problems. Next, we propose a design for sophisticated periodic execution control in order to solve these problems. Finally, we discuss the realization of periodic execution control, the effect of this control, and the result of the evaluation.

Keywords: Periodic execution control, scheduling, ART-Linux, control overhead, real-time process.

1 Introduction

In embedded systems, the types of processings to be executed are limited, and many processes are executed periodically. An example of such a process is a process to control the motor of a robot that has to be executed every 5 ms. Such a process has to be executed before a deadline; this is an example of real-time processing. In this case, we need to reduce the overhead of the periodic execution control and the dispersion of the processing time.

Rate-monotonic (RM) scheduling [1] and earliest-deadline-first (EDF) scheduling are scheduling algorithms used for realizing real-time processing. The latter is an optimal scheduling algorithm for a single-processor system, but the control overhead of this algorithm is considerable. On the other hand, the control overhead of RM scheduling is small, but the algorithm cannot utilize the system completely.

ART-Linux [2][3] has been proposed as a conventional real-time operating system that can be used in devices such as robots [4]. There have been studies related to the improvement of the performance of ART-Linux [5].

In this paper, we discuss the periodic execution control of ART-Linux and clarify several problems. Next, we present a design of a sophisticated periodic execution control to solve these problems. Finally, we discuss the realization of the periodic execution control, the effect of this control, and the result of the evaluation.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 549–563, 2010.

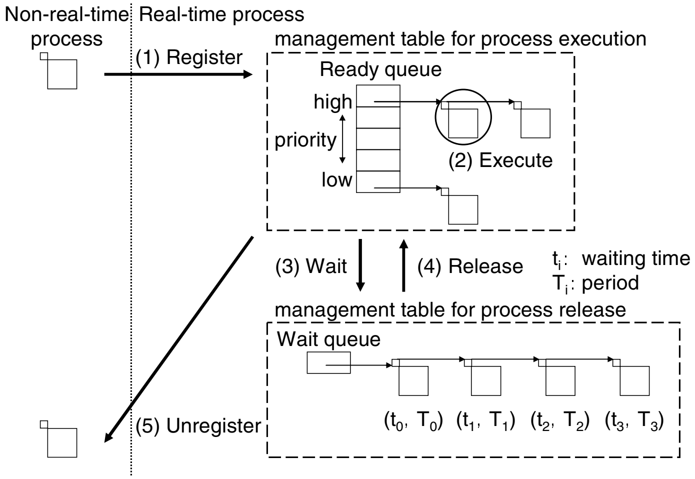


Fig. 1. Periodic execution control of ART-Linux

2 ART-Linux

2.1 Periodic Execution Control

ART-Linux is a real-time operating system based on Linux. We refer to processes executed periodically as real-time processes; all other processes are termed non-real-time processes.

ART-Linux manages real-time processes and non-real-time processes separately. Real-time processes have a higher priority than non-real-time processes. That is, non-real-time processes are executed only when there are no real-time processes to be executed.

We show the periodic execution control of ART-Linux in figure 1. The ready queue in figure 1 manages the real-time processes that are in a ready state so that the processes can be executed on the basis of priority. We call this the ready state. The wait queue manages the real-time processes that are in a wait state. We call the transition from the wait queue to the ready queue a process release. In the next few paragraphs, we explain the processing of the periodic execution control of ART-Linux.

1. Register
   When a non-real-time process requires registration, this process is connected to the ready queue. Consequently, it is managed as a real-time process. This real-time process is assigned a period and a priority.
2. Execute
   If there is already a real-time process in the ready queue, the real-time processes are executed on the basis of their priorities.
3. Wait
   A real-time process requires a wait when the processing of the real-time process for one period is finished. Then, this real-time process is disconnected from the

Table 1. Measurement environment

CPU	Pentium II 400 MHz
Memory	96 MB
Timer interrupt period	1 ms
Connection	None

ready queue, connected to the wait queue, and set a waiting time. The waiting time is the time from the next release time of the real-time process before this one in the wait queue to the next release time of this real-time process. In this way, the real-time process waits until its next release time.

4. Release

   At the time of a timer interrupt, the waiting time of the real-time process at the top of the wait queue is calculated. If the resulting value is less than 0, the real-time process is disconnected from the wait queue and connected to the ready queue. Then, the waiting time of the next real-time process in the wait queue is calculated, and the process release of this real-time process is judged. The above step is repeated until the waiting time after calculation is greater than 0 or until all real-time processes are disconnected from the wait queue. In other words, the above step is repeated until all real-time processes reaching a release time are connected to the ready queue.

5. Unregister

   When a real-time process requires unregistration, this process is disconnected from the ready queue and managed as a non-real-time process. Then, its period and priority are initialized.

2.2 Performance

2.2.1 Measurement Environment and Items

We measured the processing time from the timer interrupt to the execution of a real-time process to clarify the periodic execution control overhead in ART-Linux. Further, we measured the interval error of the periodic execution to clarify the precision of the periodic execution control. An interval of the periodic execution is the time from a processing start time to the next processing start time for a real-time process. In addition, the interval error is the difference in value between an interval and the period of a real-time process. We list the measurement environment in table 1. We used the rdtsc instruction to record the time.

2.2.2 Time from Timer Interrupt to Execution of Real-Time Process

We registered one real-time process with a period of 1 ms and the maximum priority and measured the time from the timer interrupt to the execution of this process 100 times. This result is shown in table 2, and the distribution of the time is shown in figure 2.

The difference between the average values and the maximum values is considerably larger than the difference between the average values and the minimum values in both table 2 and figure 2. We speculate that the dispersion of the time from the timer

Table 2. Time from timer interrupt to execution of real-time process

Maximum (μs)	5.47
Average (μs)	2.84
Minimum (μs)	2.57
Dispersion (μs^2)	0.12

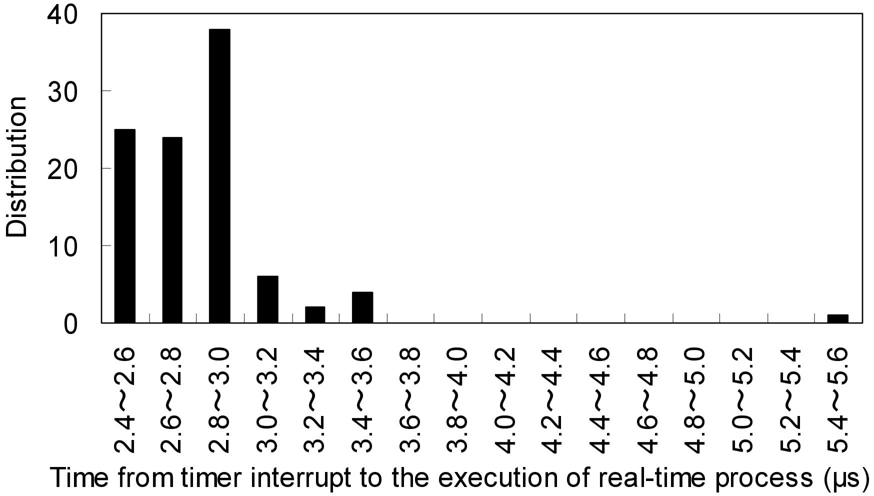


Fig. 2. Distribution of time from timer interrupt to execution of real-time process

interrupt to the execution of the real-time process is large. One reason for this may be the fact that the processing time for the process release is proportional to the number of real-time processes to be released at the same time, and the time from the timer interrupt to the execution of the real-time process increases. Therefore, we believe that the precision of the periodic execution will be low.

2.2.3 Measurement of Interval of Periodic Execution

We registered one real-time process with a maximum priority and measured the interval of periodic execution. We set the period of the real-time process to be 1, 2, 4, 5, or 10 ms and measured the interval of periodic execution in each case. The interval error is shown in figure 3; we have explained it below.

The interval error of the periodic execution is proportional to the period shown in figure 3. We measured the precision of the timer interrupt period to clarify this factor. As a result of the measurement, we found that the actual timer interrupt interval is approximately 0.996 ms for the timer interrupt interval of 1 ms. Therefore, there is approximately 0.4% error for the timer interrupt interval. In figure 3, the interval error of the periodic execution is approximately 4 μs for a period of 1 ms. That is, this error is affected by the error of timer interrupt interval. In addition, we speculate that the timer interrupt does not occur by a correct period because of the difference between the timer interrupt period that the system requires and the precision of the timer movement clock.

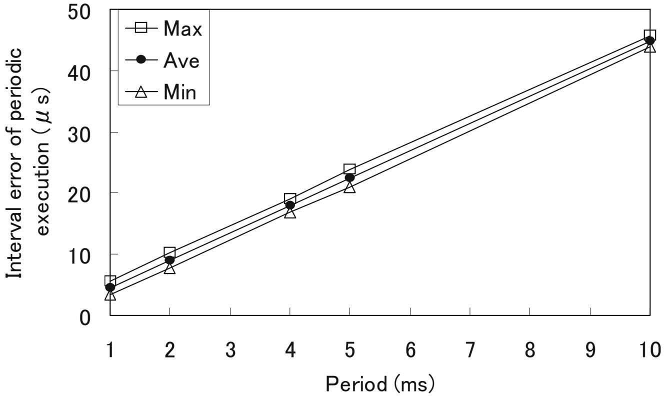


Fig. 3. Interval error of periodic execution (ART-Linux)

2.3 Problems

We will discuss five problems to be clarified by the processing of the periodic execution control and the result of the measurement in ART-Linux.

1. The control overhead for a process release is proportional to the number of real-time processes scheduled to be released at the same time, and the dispersion of the processing time for the process release is considerable.

   For the process release at the time of timer interrupt in ART-Linux, the waiting time for a real-time process at the top of the wait queue is calculated, and the release of this process is judged. Then, all real-time processes scheduled to be released at this time are disconnected from the wait queue and connected to the ready queue. Therefore, the control overhead for the process release is proportional to the number of real-time processes scheduled to be released at the same time. Therefore, if there are a considerable number of real-time processes scheduled to be released at the time of the timer interrupt, the processing time for the process release will be significant. Further, when the dispersion of the number of real-time processes scheduled to be released at each time is considerable, the dispersion of the processing time is significant.
2. The control overhead of the connection processing to the wait queue for the wait is proportional to the number of processes, and the dispersion of the processing time for the wait is considerable.

   For the wait of a real-time process in ART-Linux, the wait queue is searched from the top to calculate the waiting time of a real-time process that requires a wait and decide a position to insert this process into the wait queue. Therefore, processing that is proportional to the length of the wait queue is required; the control overhead for the wait is affected by the number of processes and the waiting time of each real-time process in the wait queue. For example, if there are a considerable number of real-time processes that are released earlier than the real-time process that requires a wait, the processing time for the wait is significant. Hence, the processing time for the wait is long, and the dispersion of this processing time is considerable.

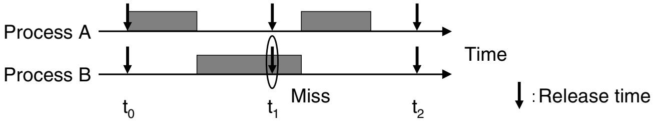


Fig. 4. Example of real-time process that is not completely processed before next release time

3. A real-time process that is not completely processed before the next release time has an adverse effect on the other real-time processes.

   Let us consider a case in which a real-time process is not completely processed before the next release time of a periodic execution control. Such a real-time process changes the interval of the periodic execution of the other real-time processes, and the precision of the periodic execution becomes low. In figure 4, process B is an example of such a real-time process. When process B requires a wait in ART-Linux, it waits from this time to the next release time. In this case, process B waits until the release time of t_2 not t_1. In other words, even if process B is not executed periodically on the basis of its period, the execution of this process continues periodically. Process B uses the time when process A is feasible, and the interval of the periodic execution of process A changes.
4. The dispersion of time from the timer interrupt to the execution of a real-time process is significant.

   As shown in Section 2.2.2, there is a case in which the change in the time from the timer interrupt to the execution of a real-time process is considerable. If such a change in time is big, the precision of the periodic execution is low. Further, the dispersion of the processing time for the process release is significant.
5. A difference in the clock precision between a timer and a processor is not coordinated.

   As shown in Section 2.2.3, the interval error of the periodic execution is proportional to the period and increases because there is a difference in the clock precision between a timer and a processor. If the difference between the period and the interval of the period execution is considerable, the precision of the periodic execution will be low.

3 Sophisticated Periodic Execution Control

3.1 Design

The processing time of the process release and the wait for the periodic execution control of ART-Linux are affected by the number of real-time processes, and the dispersion of those processing time is significant. This results from the waiting time of a real-time process. Therefore, the sophisticated periodic execution control that we propose manages the real-time process scheduled to be released with respect to the time of each timer interrupt without managing the real-time processes that need to wait in one queue. Therefore, the proposed control can solve four of the five above-mentioned problems for the periodic execution control in ART-Linux.

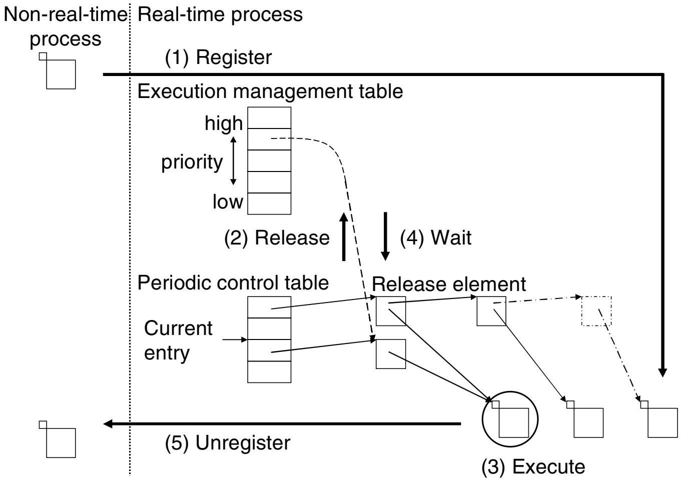


Fig. 5. Sophisticated periodic execution control

In figure 5, we show the schematic representation of the sophisticated periodic execution control that can solve problems 1–4 of ART-Linux. The control can be explained as follows: The execution management table in figure 5 is equivalent to the ready queue shown in figure 1, and the periodic control table is equivalent to the wait queue.

The execution management table manages the real-time process to be executed. In particular, it manages the release element of the real-time process to be executed by connecting it to the entry of the execution management table on the basis of the priority of the real-time process. The release element is the element that has the information of the real-time process and has two queue entries. These entries are used for connecting the real-time process to the execution management table and the periodic control table. The periodic control table manages the real-time process scheduled to be released each time; the number of entries is equivalent to the longest period that can be set. One entry in this table is equivalent to one timer interrupt and is connected to the release element of the real-time process to be released at the timer interrupt. In other words, the number of entries in the periodic control table should be the least common multiple of the periods for all real-time processes.

We explain the processing of the sophisticated periodic execution control below.

1. Register
   When a non-real-time process requires registration, this process is managed as a real-time process by the periodic control table. Then, the periodic control table is searched, an entry that is connected to few release elements in the periodic control table is found, and the release element of the real-time process is connected to this entry. Simultaneously, the real-time process is set a period and a priority.
2. Release
   At the time of the timer interrupt, the current entry is translated to the next entry. The current entry is the entry that corresponds to the time of the timer interrupt in the periodic control table. If there is a release element in the current entry, the

decision to release a real-time process is taken. In particular, the release element is connected to the execution management table on the basis of the priority of the real-time process. Then, the release element is not disconnected from the periodic control table. If the priority of the real-time process to be released is greater than that of the current process, a pre-emption occurs. In addition, if there is a release element of the real-time process to be executed in the execution management table at this time, all release elements of this real-time process are disconnected from the execution management table and the periodic control table, and this real-time process is terminated. This is done to restrain the adverse effect that this real-time process has on the other real-time processes.

3. Execute
   After the processing of release, wait, and unregister, if there is a release element in the execution management table, real-time processes are executed on the basis of their priorities. Then, the release element is not disconnected from the execution management table. Therefore, there is a release element of the real-time process being executed at the top of the entry corresponding to the priority of the execution management table. A bitmap is used for deciding which entry should be connected to the execution management table.
4. Wait
   When a real-time process requires a wait, its release element is disconnected from the periodic control table. Therefore, the process waits until the next release time.
5. Unregister
   When a real-time process requires unregistration, all release elements of this process are disconnected from the periodic control table. Then, the period and the priority of this process are initialized. Hence, this process is managed as a non-real-time process.

3.2 Characteristics

The sophisticated periodic execution control that we proposed in Section 3.1 has the following characteristics:

1. The release of the real-time process is managed by the periodic control table.
2. A release element has two queue entries that are used for connecting the release element to the execution management table and the periodic control table.
3. A real-time process that is not completely processed before the next release time is terminated.
4. When the release element is connected, the dispersion for the number of release elements connected to each entry of the periodic control table becomes low.

According to characteristic 1, the calculation of the waiting time is unnecessary. Because of characteristic 2, the processing of the disconnection from the periodic control table for the release and the processing of the connection to the periodic control table for wait are unnecessary. Furthermore, because of characteristic 1 and 2, the number of queue operations required decreases, and the control overheads for the process release and wait become small. In addition, we have speculated that the dispersion of the processing time for the process release and wait become small because the number of processing for searching the release element is few. Because of characteristic 3, the influence of a real-time process that is not completely processed before the next release time on the other real-time processes is restrained. Because of characteristic 4, the dispersion of the processing time for the process release becomes small.

3.3 Realization

3.3.1 Data Structure

We have realized the sophisticated periodic execution control that we proposed earlier. We show the data structure of this periodic execution control in figure 6 and explain each management table below. The data structure of the periodic execution control has six management tables.

The release management table manages the information required to decide the release of real-time processes with the periodic control table. The current entry (curpointp) is translated into the next entry at every timer interrupt. Pct is the pointer to the top of the periodic control table, and pctsize is the number of entries in the periodic control table.

The number of entries in the periodic control table is the least common multiple of the periods of all real-time processes. Each entry manages the release element. All release elements that are connected with the entry to which curpointp points, (current entry) are connected to the execution control table for the process release. The count of the periodic control table is the number of release elements to manage.

The release element contains information on the real-time process. Next is used for connecting the release element to the periodic control table, whereas prinext is used for connecting the release element to the execution management table. Procp is pointer to the real-time process.

The execution management table manages the release elements of the real-time processes that are to be executed on the basis of their priority. In this table, the release element of the real-time process being executed is at top of the entry that has the maximum priority among the entries that release elements are connected.

The real-time information table manages the period and the priority of the real-time process. The number of each entry is equivalent to the process identifier. If the process management table manages the period and the priority of the real-time process, the access time may increase because the process management table has a considerable amount of information. Therefore, the cache miss is restrained, and the access time required to use the real-time information table becomes short. When a real-time process is registered, rflag is set a period and rpri is set a priority. If the real-time process is released, the most significant bit of rflag is set. Further, when the real-time process requires a wait, this bit is reset.

The release element management table manages the unused release elements. Num is the number of unused release elements, and maxnum is the number of all release elements that are allocated at the time of the initialization of all management tables.

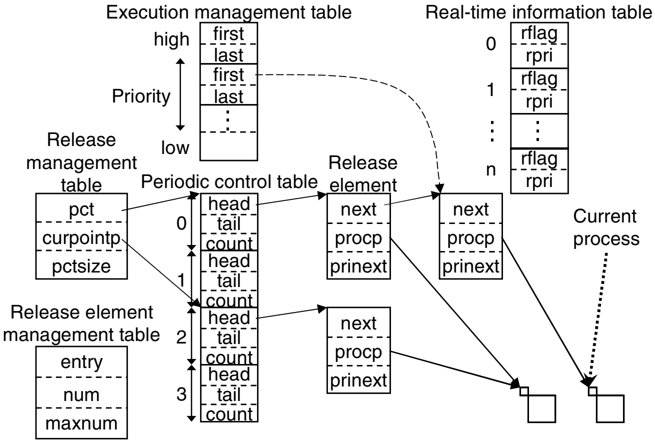


Fig. 6. Data structure

3.3.2 Coding

Bear in mind the following for the implementation of the sophisticated periodic execution control.

First, the optimization of the code and the use efficiency improvement of the register used in the method are required in order to speed up processing. For example, we summarize the cords handling the same variable in one place in order to reduce the amount of rewriting in the register. As a result, the processing time becomes short.

Many release elements may be used by a real-time process in proposed control. Therefore, the quantity of data to be referred to during one processing may become considerable, and the cache miss rate may become high. Moreover, the access time may increase significantly because the data spatial locality becomes low. Therefore, when the sophisticated periodic execution control is implemented, it is necessary to devise how to handle data such as the period and the priority of the real-time processes and how to allocate the release element.

Hence, we reduce the quantity of data for a real-time process and minimize the quantity of data for a release element. Further, during the processing of the process release and the wait, data that do not relate with the periodic control are not read. The release element has only two queue entries and a pointer to the process. Information related to the real-time process, such as the period and the priority, is described in the real-time information table. This is done to avoid giving the information of the real-time process in the process management table and to reduce the quantity of data that is accessed. If the process management table contains the information of the real-time process, the quantity of data that needs to be accessed increases considerably because the process management table has a significant amount of information. Moreover, the state of the real-time process as ready or wait is contained in rflag. We can use the same entry to manage the period and the state of the real-time process because the period of the real-time process is only changed in the processing of registration or unregistration. Therefore, we can reduce the quantity of data accessed, and the cache miss rate becomes low.

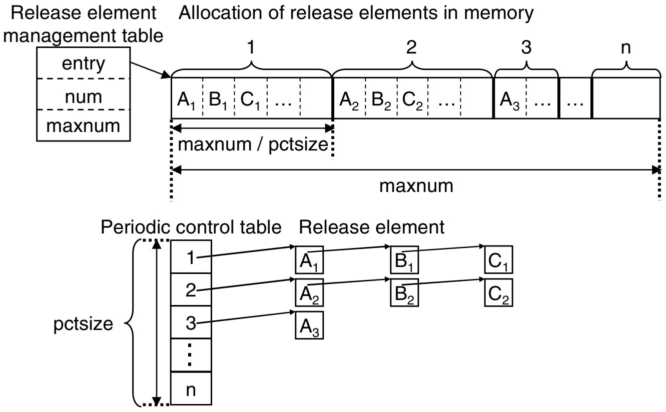


Fig. 7. Allocation of release elements

Figure 7 shows how to allocate release elements in the control that we have realized. At the initialization of each management table, release elements are allocated according to the maxnum size. Let x be the value that is obtained by dividing maxnum by pctsize. “i” is the serial number of each entry in the periodic control table. When the release element is connected to the i-th entry, the unused release element to use is searched from the (x * i)-th entry of all release elements. Therefore, the release element for each entry in the periodic control table is accumulated in the memory. Therefore, the data spatial locality is improved, and the access time for the process release and wait decreases.

3.4 Effects

We have described the effects of the sophisticated periodic execution control below.

1. The control overhead of the process release becomes small.

   For the process release of a real-time process in ART-Linux, the waiting time for all real-time processes scheduled to be released at the same time is calculated; then, these processes are disconnected from the wait queue and connected to the ready queue. On the other hand, in proposed control, the current entry is translated into the next entry, and all release elements of the current entry are connected to the execution management table. Therefore, it is unnecessary to calculate the waiting time and to disconnect the real-time processes from the wait queue. Moreover, the number of processing that is influenced by the number of real-time processes scheduled to be released at the same time decreases, and the dispersion of the processing time for process release becomes small.
2. The control overhead of the wait becomes small, and the dispersion of the processing time for the wait becomes small.

   For the wait of a real-time process, in ART-Linux, the wait queue is searched from the top, and the waiting time for the real-time process that requires a wait is calculated. Then, the process is disconnected from the ready queue and connected to the wait queue. On the other hand, in the proposed control, the release element is only disconnected from the execution management table. In addition, the release element of the real-time process being executed is at the top of the entry with its priority. Therefore, the processing time for the wait is short, and this value is fixed without affecting the number of the real-time processes in the wait queue and the waiting time of each real-time process.

3. The adverse effect of a real-time process that could not be completely processed before the next release time on the other real-time processes is restrained.

   In ART-Linux, the execution of the real-time process continues periodically even if the execution is not completed before the next release time. On the other hand, in the proposed control, if there is a release element of the real-time process still to be executed in the current entry as a process release at the next release time, this real-time process is terminated. In other words, a real-time process that is not completely processed before the next release time is terminated. Therefore, the influence of such a real-time process on the other real-time processes is restrained.

4. The dispersion of the processing time for the process release is small.

   In the proposed control, for registration, when the release element is connected to the periodic control table, the periodic control table is searched, an entry managing few release elements in the periodic control table is found, and the release element of the real-time process is connected to this entry. Therefore, the dispersion for the number of release elements being connected to each entry of the periodic control table becomes low. Hence, when there are real-time processes with different periods, the dispersion of the processing time for a process release becomes small. As a result, the dispersion of the time from the timer interrupt to the execution of a real-time process becomes small.

4 Evaluation

We have implemented the proposed control in ***AnT*** operating system [6][7]. We evaluated the periodic execution control for ***AnT*** operating system and ART-Linux in table 1. In addition, a non-real-time process that reads data of 1M byte is executed to remove influence of the CPU cache. The number of real-time processes registered in ***AnT*** operating system is the same as ART-Linux. We explain two evaluation items below.

First, to evaluate the control overhead of a process release, we measured the processing time of the process release when the number of real-time processes is N. We registered N processes with a period of 1 ms and measured the processing time of the process release at the time of each timer interrupt. In other words, we measured the processing time of the process release when the number of real-time processes scheduled to be released at the same time is N. We changed the number of real-time processes in a range from 1 to 100 and measured the processing time in each case. The result is shown in figure 8.

Second, to evaluate the control overhead of the wait, we measured the processing time of the wait. We registered N processes with a period of 1 ms, and measured the processing time of the wait for the N-th real-time process executed after timer interrupt. We changed the number of real-time processes in a range from 1 to 100 and measured the processing time in each case. The result is shown in figure 9.

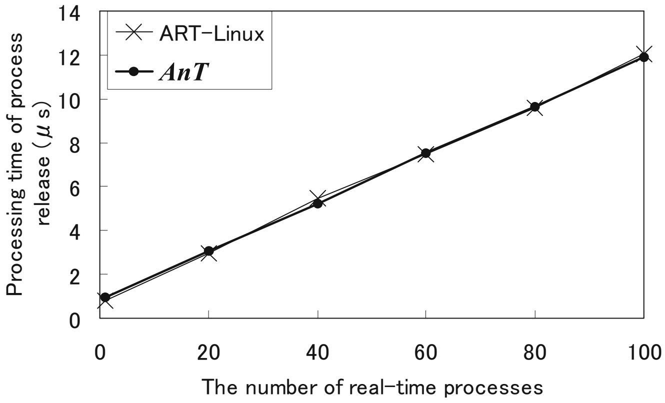


Fig. 8. Processing time of process release

In figure 8, the processing time for ***AnT*** operating system is equal with that for ART-Linux. In ART-Linux, for the process release of a real-time process, the waiting time for all real-time processes scheduled to be released at the same time is calculated; then, these processes are disconnected from the wait queue and connected to the ready queue. On the other hand, in proposed control, the current entry is translated into the next entry, and all release elements of the current entry are connected to the execution management table. Therefore, it is unnecessary to calculate the waiting time and to disconnect the real-time processes from the wait queue. We speculated that the processing time of the process release becomes short, but the results of ***AnT*** operating system and ART-Linux do not have the difference in figure 8. This reason may be the fact that the data spatial locality of ***AnT*** operating system is lower than that of ART-Linux.

In figure 9, the processing time for ***AnT*** operating system is short and this value is fixed. The processing time for ART-Linux is proportional to the number of real-time processes in figure 9. On the other hand, the value of the processing time for ***AnT*** operating system is fixed and small without affecting the number of the real-time processes. In ART-Linux, for the wait of a real-time process, the wait queue is searched from the top, and the waiting time for the real-time process that requires a wait is calculated. Then, the process is disconnected from the ready queue and connected to the wait queue. On the other hand, in the proposed control, the release element is only disconnected from the execution management table. Therefore, the processing time of the wait for ***AnT*** operating system is short and this value is fixed.

In addition, for periodic execution control, (N - 1) number of real-time processes require a wait until the N-th real-time process is executed after timer interrupt. Therefore, the processing time of the wait until the N-th real-time process is executed is the sum total of that for 1st–(N - 1)-th real-time processes. In other words, for the proposed control, the processing time until the 100th real-time process is shortened about 50μs. This value is about 5% for a period of 1 ms.

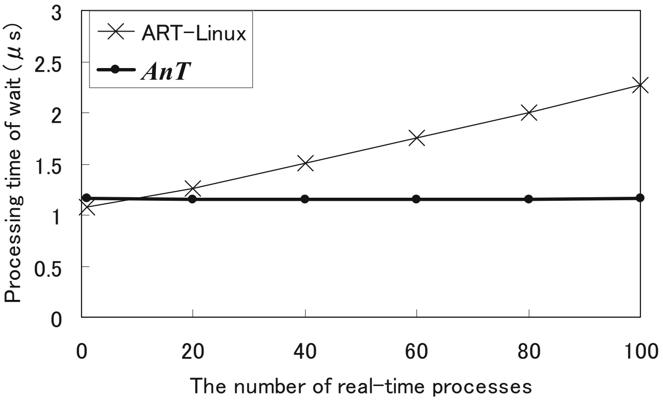


Fig. 9. Processing time of wait

5 Conclusion

We clarified several problems in ART-Linux and proposed a sophisticated periodic execution control to solve these problems.

The problems in ART-Linux are as follows: (1) The control overhead for a process release is proportional to the number of real-time processes scheduled to be released at the same time, and the dispersion of the processing time for the process release is considerable. (2) The control overhead of the connection processing to the wait queue for a wait is proportional to the number of processes, and the dispersion of the processing time for the wait is considerable. (3) A real-time process that is not completely processed before the next release time has an adverse effect on the other real-time processes. (4) The dispersion of the time from the timer interrupt to the execution of a real-time process is significant. (5) The difference in the clock precision between a timer and a processor is not coordinated.

We proposed a method to manage the real-time process scheduled to be released at the time of each timer interrupt as a solution to four of the above-mentioned problems. When the method is used, the control overhead for the process release and the wait decrease, and the dispersion of the processing time for them becomes small. In addition, the adverse effect of a real-time process that was not completely processed before the next release time on the other real-time processes was reduced. We have realized the proposed control and explained about the date structure. In addition, we discussed the points to bear in mind while coding. Moreover, we evaluated our suggestion.

Our future work will include survival evaluations of our suggestion and the development of a method to solve survival problems.

Acknowledgement. This research was partially supported by Grant-in-Aid for Scientific Research 21500055.

References

1. Liu, C., Layland, J.: Scheduling algorithms for multiprogramming in a hard real-time environment. Journal of the ACM 20, 46–61 (1973)
2. Ishiwata, Y.: SMP kernel based stabilization of ART-Linux and measurement of its real-time processing performance. In: SI 2002 (2002)
3. Ishiwata, Y.: Development of ART-Linux on SH-4 processor and its application to quality control. In: SI 2002 (2002)
4. Yokoi, K., Kanehiro, F., Kaneko, K., Kajita, S., Fujiwara, K., Hirukawa, H.: Experimental Study of Humanoid Robot HRP-1S. Intl. J. Robotics Research 23(4-5), 351–362 (2004)
5. Hori, Y., Nakajima, T., Katashita, T., Sekiyama, M., Toda, K.: Approaches to Improving Performance of ART-Linux with Dedicated Hardware. IPSJ Technical Reports, 2004-SLDM-119, vol. 2005(27), pp.109–114 (2005)
6. Taniguchi, H., Nomura, Y., Tabata, T., Adachi, T., Nomura, Y., Umemoto, M., Nishina, T.: Design for AnT Operating System. IPSJ Technical Reports, 2006-OS-103, vol. 2006(86), pp. 71–78 (2006)
7. AnT operating system, `http://www.swlab.cs.okayama- u.ac.jp/lab/tani/research/AnT/index.html`

Videogames and Virtual Reality as Effective Edutainment Tools

Akemi Gálvez and Andrés Iglesias

Department of Applied Mathematics and Computational Sciences,
University of Cantabria, Avda. de los Castros,
s/n, E-39005, Santander, Spain
{galveza,iglesias}@unican.es
http://personales.unican.es/iglesias

Abstract. This paper describes a new edutainment-based teaching approach aimed at fostering students' engagement in the classroom at university and college levels. The approach, designed to comply with Bologna declaration's regulations and requirements, relies on the intensive use of multimedia computer tools, notably videogames and virtual reality technology, as the appropriate means to motivate students and make the courses amusing and engaging. The paper presents and analyzes a case study: the application of our approach to an introductory Computer Graphics course for junior Computer Science students at the University of Cantabria, discussing some important educational issues related to our approach and showing several examples of typical projects carried out during the course.

1 Introduction

This paper is strongly motivated by the ongoing efforts carried out in most European universities and other higher education institutions in order to adapt our teaching and learning structures, systems, methods and tools to the upcoming European Higher Education Area (EHEA). This process, initiated with Bologna declaration (signed in 1999 by 29 European countries and seen today as the well-known synonym for the whole process of reformation in the area of higher education), aims at creating *a European space for higher education in order to enhance the employability and mobility of citizens and to increase the international competitiveness of European higher education* [3,9,12]. The upmost goal of this process is the commitment freely taken by each signatory country to reform its own higher education systems in order to create overall convergence at European level. This process encompasses the adoption of a common framework of readable and comparable degrees as well as the introduction of undergraduate and postgraduate levels in all countries along with ECTS (European Credit Transfer System) credit systems to ensure a smooth transition from one country's system to another one, thus enforcing free mobility of students, teachers and administrators among the European countries.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 564–576, 2010.

According to EHEA calendar, this harmonization process must be ready to start in 2010. Although many advances have already been done, developments focused especially on academic aspects, such as the definition of the new curricula and grading systems. Amazingly enough, little attention has been placed so far upon the definition of suitable strategies for teaching and learning. This fact is especially notorious in college and university levels [1]. Fortunately, there is a wealth of educational methods that can be effectively applied to fulfill Bologna declaration's principles and regulations. Among them, *edutainment* (i.e. the integration of education and entertainment) has been often referred to as a highly-efficient approach for effective learning [2,11,14]. By using current educational tools at full extent, learning becomes fun and teachers instruct students in a manner which is both engaging and amusing. This is the approach we follow in this paper.

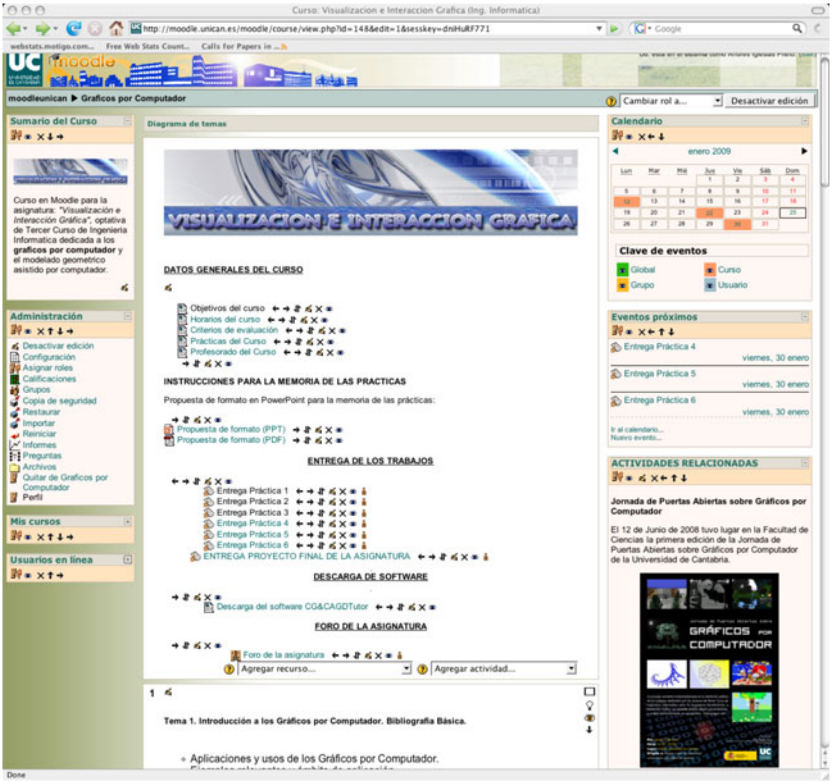
Videogames and virtual reality (VR) have been often mentioned as technological tools with the power to facilitate learning. Videogames are a great way to encourage students to be interactive with the subject being learned. They are also an excellent way to make learning not feel like work. At its turn, virtual learning environments provide rich teaching patterns and teaching contents [14]. They are also helpful in order to improve students' skills to analyze problems and explore new concepts in a way that would otherwise not be possible.

A very important (and not very well explored yet) issue is the role that videogames and virtual reality play in powering students' engagement. Educational systems strongly rely on students engagement as it provides the opportunity to teach under different learning styles and skill levels. Our claim in this paper is that videogames and VR are essential tools (arguably the best ones) for powering students engagement and getting the best of them in the topics under study. Such a claim will be supported by the presentation and analysis of a case study, as explained in next sections.

The structure of this paper is as follows: in Section 2 we describe some educational issues regarding our edutainment approach, such as learning resources, teaching methods and workflow. Then, Section 3 reports a case study: the application of our approach to an introductory Computer Graphics course at the University of Cantabria. We describe the subject and students' profile and give some examples of typical course projects. The paper closes with the conclusions and some further remarks.

2 Educational Issues

In this section we discuss the learning resources we use and teaching methods we apply in our course and describe how videogames and VR have been used in order to promote our students' engagement at the classroom and to effectively learn difficult concepts and methods.



Curso en Moodle para la asignatura: *"Visualización e Interacción Gráfica"*, optativa de Tercer Curso de Ingenieria Informatica dedicada a los **graficos por computador** y el modelado geometrico asistido por computador.

DATOS GENERALES DEL CURSO

- Objetivos del curso
- Horarios del curso
- Criterios de evaluación
- Prácticas del Curso
- Profesorado del Curso

INSTRUCCIONES PARA LA MEMORIA DE LAS PRACTICAS

Propuesta de formato en PowerPoint para la memoria de las prácticas:

- Propuesta de formato (PPT)
- Propuesta de formato (PDF)

ENTREGA DE LOS TRABAJOS

- Entrega Práctica 1
- Entrega Práctica 2
- Entrega Práctica 3
- Entrega Práctica 4
- Entrega Práctica 5
- Entrega Práctica 6
- ENTREGA PROYECTO FINAL DE LA ASIGNATURA

DESCARGA DE SOFTWARE

- Descarga del software CG&CAGDTutor

FORO DE LA ASIGNATURA

- Foro de la asignatura

1

Tema 1. Introducción a los Gráficos por Computador. Bibliografía Básica.

- Aplicaciones y usos de los Gráficos por Computador.

enero 2009

Lun	Mar	Mié	Jue	Vie	Sáb	Dom
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

Clave de eventos

🟩 Global	🟧 Curso
🟨 Grupo	▨ Usuario

Eventos próximos

Entrega Práctica 4
viernes, 30 enero

Entrega Práctica 5
viernes, 30 enero

Entrega Práctica 6
viernes, 30 enero

Ir al calendario...
Nuevo evento...

ACTIVIDADES RELACIONADAS

Jornada de Puertas Abiertas sobre Gráficos por Computador

El 12 de Junio de 2008 tuvo lugar en la Facultad de Ciencias la primera edición de la Jornada de Puertas Abiertas sobre Gráficos por Computador de la Universidad de Cantabria.

Fig. 1. Moodle platform of the course

2.1 Learning Resources

One fundamental ingredient for engagement is motivation. In our course, we used almost any imaginable media to capture students attention and curiosity:

- **TV screen captures:** An arguably advantage of this media is that we all are used to watch TV since our early childhood and to learn by playing, either solving a puzzle, or singing along or whatever else. Because of that, we have used them intensively to illustrate students about different concepts, as in Example 1 described later.
- **guest speakers:** We invited people to give talks to our students about a particular topic. For instance, in order to create the graphical output shown in Section 3, students were given an Open GL introductory course by Prof. F. Luengo (Univ. of Zulia, Venezuela). Another colleague (Enrique Bernardos) with large experience in flight simulators gave us a smooth introduction to that topic. As a general procedure, we invite everyone with any useful knowledge to offer to our students, and the experience has been very positive in all cases.

- **books, journals, videotapes, freeware:** Selected bibliography is provided to help students in their assignements. All bibliographic entries are available in our library and are presented and discussed at the classroom during the first week of the course so students get a high degree of autonomy regarding the information sources. Similarly, we use freeware exclusively in all our projects, thus keeping student and institutional loads to the minimum [6,7,8,10,13]. We also give students the freedom to explore by themselves; they can not only work on the project they like the most, but also choose the computer tools and methods to be used in that project. We think that, even if their choice reveals eventually wrong, it will be productive once students analyze why it failed and how to improve it for the next time.
- **Internet:** Since it is one of the most effective and engaging ways to acquire information, students are encouraged (even required) to surf at the Web looking for material and exploring new ways to acquire information, download free software and contents (manuals, examples, etc.) and as an effective communication channel. Simultaneously, we created Internet tools for all steps of the learning process. The course is available on *Moodle* (see Figure 1) to offer our students a free and open source platform for rich interactions. To this aim, *Moodle* functionalities are used at full extent, from the glossary of terms for any new chapter treated at the classroom to a chat where students and teachers can share their impressions about the course, make questions and get the answers, a repository of material for classes and projects, a private space at the server where students can upload their assignments, an authentication system to check for access control, a management system to assign roles according to user's profile (system administrator, course manager, teacher, student, guest), quizzes and quiz questions, calendar of activities, syndication using RSS, etc. Finally, private inter-group communication was mostly performed by using social networks such as *Facebook*, *MySpace*, *Twitter* and *Tuenti*. Many students also had the initiative to publish trailers of their projects in *YouTube* and other video sharing websites.
- **Others:** This category comprises videogames, virtual reality scenes and worlds, public presentations and discussions at the classroom, collaborative work, etc. Since our approach mostly rely on these materials and methods, they are discussed in detail in next sections.

2.2 Teaching Methods

A clear advantage of using computer software for educational purposes is that it emphasizes thinking. In this sense, videogames and virtual reality scenarios are excellent candidates (arguably the best ones) for computer edutainment. However, this approach is not well suited for traditional teaching methods, and hence new methods are to be applied instead. In our courses we combine traditional teaching methods with other different techniques: scaffolding, project-based learning and collaborative problem solving.

Scaffolding aims at providing early support for learning whenever new concepts and skills are being first introduced. Such supporting assistance is gradually removed as students develop their own strategies for learning. Combined with traditional teaching, it proved to be a powerful approach at the initial stages of the learning process. In our courses, it is mostly used for learning a new computer program or at the beginning of a new project so that students get the required materials (information sources, templates and guides, compelling tasks) from the very beginning allowing them to become autonomous afterwards.

In *project-based learning* we pursue students play the role of teachers. Once a new problem is introduced, students are requested to collect data and information, analyze it, design the steps of the project, debate them with group members, and draw conclusions. Projects described in next section are all examples of this kind of learning approach. Projects are freely chosen by students from a pool of optional projects. Alternatively, students are allowed to propose projects by themselves.

Finally, *collaborative learning* is very useful in order to develop students' social skills and strategies when challenged about a problem which is too much complex to be solved individually thus requiring a team of people working together. We experienced a type of collaborative learning usually referred to as *jigsaw puzzle*. Students are arranged in small groups to deal with a new idea, concept or method which, once acquired, must be taught to their classmates. Groups are flexible throughout the course in order to promote social skills and interactivity amongst students.

2.3 Workflow

In general, the workflow for a new subject starts with the presentation of the concepts to be learnt. Such a presentation is generally handled with the help of different media such as newspapers, videotapes, TV movies and others. Teachers at this stage are mostly motivators, their goal being to stimulate students' curiosity by examples. After this gentle overview, the concept is then explored in detail. Theoretical explanations in a rather traditional way are mixed up with scafffolding until the main ideas, methods and resources are acquired. Teachers now move to a new role as guiders, offering advice and support in a gradual way.

Then, students move deeper into the subject by collaborative learning. The role of teachers is now taken by students, so real teachers assume a new role as students' work supervisors. They also monitor group's evolution, correct mistakes and ensure discussions are properly driven.

In advanced stages, teachers propose projects or evaluate students' proposals for feasibility. Once projects are accepted, teachers supervise, give advices, make suggestions and finally validate and evaluate them in order to determine students' degree of comprehension about the subject under study.

During the whole process, teachers also play an important role about the motivational issues in order to keep students' motivation and creativity alive and make each lesson fun.

3 A Case Study: Computer Graphics at the University of Cantabria

This section reports a case study on the application of our computer edutainment approach to a real case. We begin our discussion by describing the subject to be taught and our students group. Then some illustrative examples of projects carried out under our approach are brielfy outlined. Finally, some motivational strategies are also reported.

3.1 Subject and Students

The edutainment approach described above has been successfully applied during the last three years to a Computer Graphics course for junior (third-year) Computer Science students at the University of Cantabria. It is a Fall-semester elective course of 7.5 credits aimed at introducing students to computer graphics techniques and their applications to scientific and technical fields. The group in Fall 2007 had 22 students while for Fall 2008 the group increased to 32, with 37 students for the current year.

As a part of their assignments, students are required to develop one or several projects for different subjects such as Geometry, Mechanics, Economics, Computer-Aided Design (CAD), Programming, Operating Systems, etc. to be subsequently used in studies ranging from Mathematics or Physics to Economics, Architecture, Computer Science, Mechanical Engineering and others. Projects are typically videogames or virtual reality scenarios (or movies) involving topics from those subjects. Expertise from every specific field is acquired through the interaction with other students from those degrees, thus promoting students' social skills and collaborative work. The final product has also been used at instructional level of students in such degrees, with also very positive results in terms of students' engagement and motivation.

3.2 Project Examples

In this section we describe some basic examples of projects developed by our students for specific courses. In agreement to our edutainment approach, our description will be mostly visual and based on videogames and/or VR examples. For each example we provide the concept and skills to be acquired, the computer tool generated (either a videogame or a VR scene or world) and a brief discussion about the generation process. All examples have been created from scratch by our students, who are credited by giving their corresponding names at the example header.

EXAMPLE 1.- **Subject:** *Geometry & Algebra, Computer Graphics.* **Concept**: *2D-Transformations, Filling Algorithms.* **Project**: *Interactive Movie.* **Authors**: *Sergio del Valle, Miguel A. Jiménez, Alvaro Fernández.* **Project**: *Chessboard movie.* **Authors**: *Fernando Aguilar, Ignacio Ara, Diego García.*

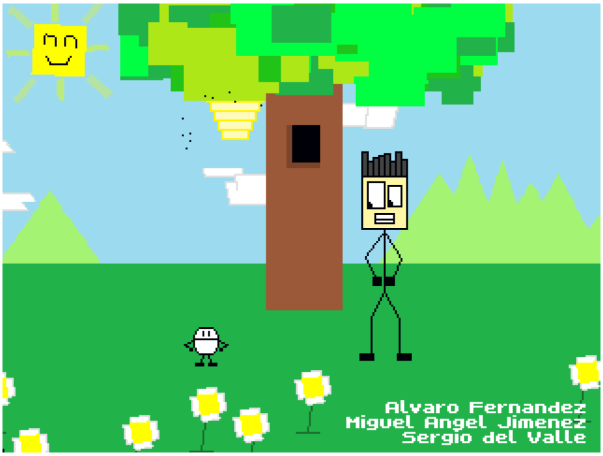


Fig. 2. Interactive movie to illustrate 2D transformations

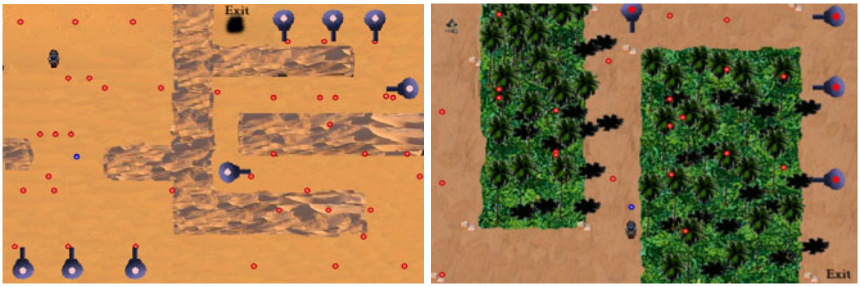


Fig. 3. Two screeenshots of“The Last Chance” videogame

In this example the goal is that students develop a geometrical insight of the two-dimensional transformations being able to understand, manipulate and modify 2D objects and create simple animations by using 2D transformations exclusively. Students are shown some examples of useful 2D transformations in real life (such as spirals, mosaics, fractals, etc.) as well as movies created with simple 2D transformations. Inspired by this input, our students developed an interactive 2D movie based on the same principles and methods (see Figure 2). To improve the graphical output, some filling algorithms to color the objects have been added. Characters in this movie can change their physical appearance and interact with users and other characters. To strengthen students' geometrical intuition, a number of actions have been included: characters can talk, move, fly, jump, etc. The movie was implemented in freeware *Fenix* on an Intel PC platform equipped with Linux.

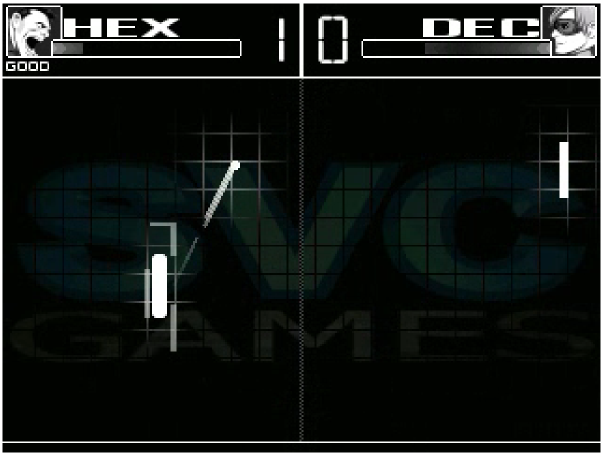


Fig. 4. "Dirty Bit X" videogame

EXAMPLE 2.- **Subjects:** *Geometry & Algebra, Programming.* **Concept**: *2D-Transformations, Collision Avoidance, Process & Thread Concurrency and Synchronization.* **Project**: *Videogame "The Last Chance".* **Authors**: *Rubén García, Javier Quintana, Ricardo Ruiz*; **Project**: *Videogame "Dirty Bit X".* **Author**: *Sergio del Valle.*

Figures 3 and 4 show screenshots of the videogames "The Last Chance" and "Dirty Bit X" respectively, designed to illustrate the concepts of 2D transformations (like Example 1) and collision avoidance, a classical topic in Robotics, Computer Animation, Geographical Information Systems and others. The first game is able to handle as many as 200 processes simultaneously, so it is also a good example of concurrency and synchronization of processes and threads, a very important subject in programming. The second game, a modification of the classical *Pong*, also include routines for movement based on the laws of Physics and has been used to explain concepts such as velocity, acceleration, force, torque, energy and others. Both games have been implemented in *Fenix* on Linux platforms.

EXAMPLE 3.- **Subjects:** *Mechanics, Mechanisms, machines & robots.* **Concept**: *Degrees of Freedom, Hamiltonian Systems.* **Project**: *VR Hydraulic water wheel.* **Authors**: *Fernando Aguilar, Ignacio Ara, Diego García.*

Figure 5 shows a screenshot of a VR animated scene of a water wheel. It was created by using the freeware modeler *Blender* for the mechanical parts, then *Javascript* for user interaction and animation. This structure is used in Physics and Mechanical Engineering degrees to analyze the intrinsic behavior of mechanisms. During the modeling stage, students become skilled in identifying the different parts of the structure (buckets, axle, tub, ring gear, etc.) by creating them by themselves. Then, the animation stage introduces students into the fundamentals of their kinematics and dynamics. Interactive experiments can

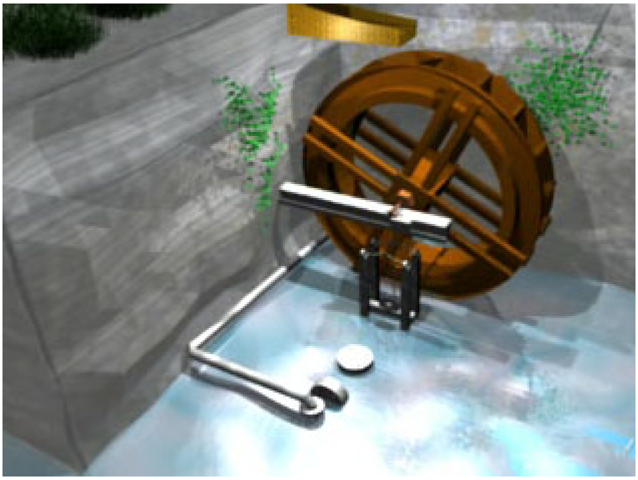


Fig. 5. VR model of a water wheel

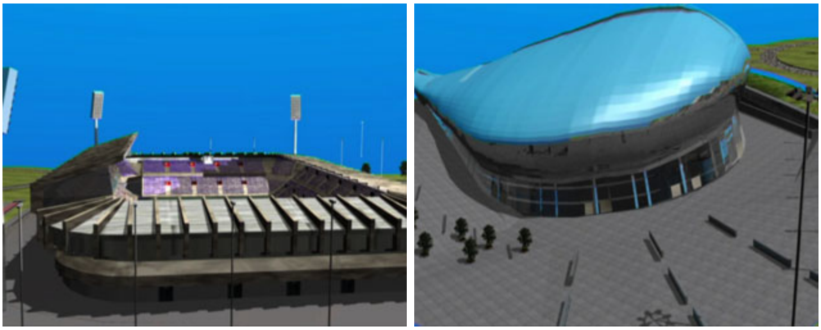


Fig. 6. (Left) Santander Football Stadium; (right) Santander Omnisports Arena

be conducted with the generated scene, ranging from immersive navigation to parameter modification in order to analyze/predict mechanism's response.

EXAMPLE 4.- **Subjects:** *Computer Design, Architecture.* **Concepts**: *Architectural Design, Fundamentals of CAD.* **Project**: *Virtual Santander.* **Authors**: *Fernando Aguilar, Ignacio Ara, Diego García.*

A virtual version of our city (Santander) has been created in *Blender* and then used to instruct students about the fundamentals of Architectural Design. The process of creating the virtual city is also useful in Civil Engineering, Mechanical Engineering and CS degrees in subjects such as CAD and computer modeling. Figure 6 shows two examples of interesting buildings of Santander: the football stadium and the Omnisports Arena, respectively.

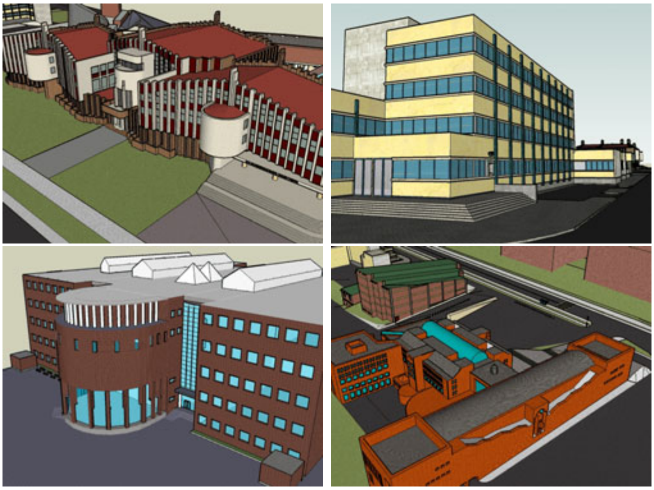


Fig. 7. Four "Virtual Campus" screenshots

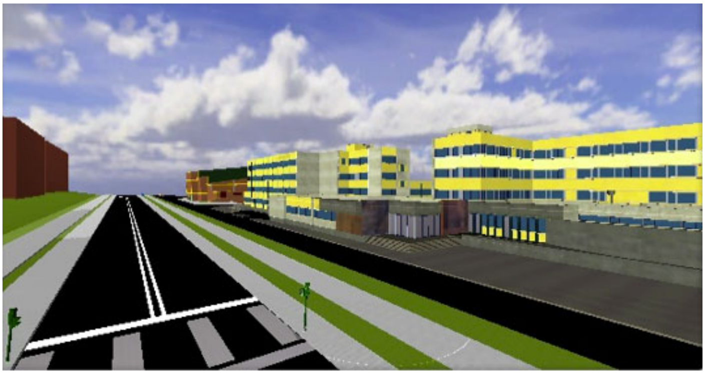


Fig. 8. Main street of "Virtual Campus"

<u>**EXAMPLE 5**</u>.- **Subjects:** *Computer Design, Architecture.* **Concepts**: *Architectural Design.* **Project**: *Virtual Campus.* **Authors**: *José A. Abascal, Ignacio Bustillo, Laro Fernández, Adrián Fernández, Fernando Martín, Manuel Pando, Angel Tezanos.*

Similarly to the previous example, a virtual campus of our University has also been generated through *Google*'s freeware *SketchUp* modeling software. Figure 7 shows four screenshots of the campus. In this case, pretty simple textures have been used in order to keep the whole model at reasonable size. Such a model has been further processed for efficient rendering and texturing through *Apple*'s software tool *Unity*. Figure 8 depictes the final image of the campus main street.

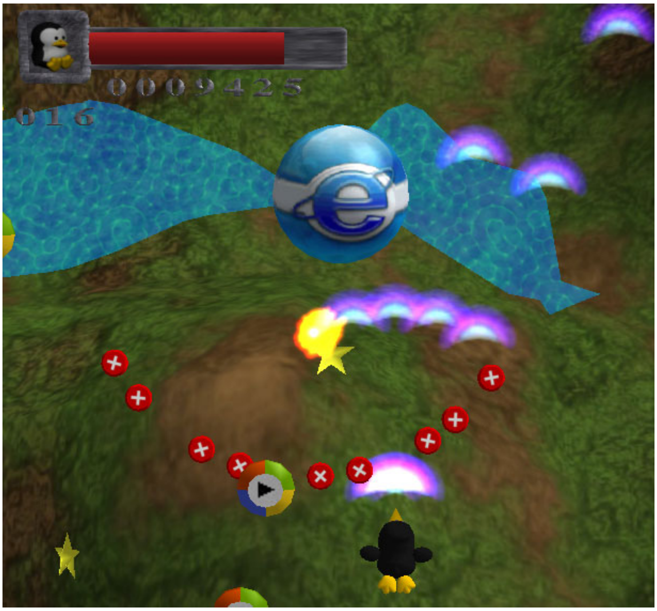


Fig. 9. "TuxInGates" videogame

EXAMPLE 6.- **Subject:** *Operating Systems.* **Concept**: *Operating System Suites & Applications.* **Project**: *Videogame "TuxInGates".* **Author**: *David Garay.*

This videogame was originally motivated by the Operating Systems (OS) course which is compulsory for our CS students. In addition to learning the in-depth fundamentals of general operating systems and UNIX, the most popular proprietary (Microsoft, MacOS) and open-source (Linux) operating systems for microcomputers are also studied. *"TuxInGates"* is a standard shoot'em up videogame where user plays the role of the classical *"Tux the Penguin"* (Linux kernel's official mascot) as a multi-directional shooter fighting Microsoft (MS) enemies represented as MS-Dos, Windows, Office, Encarta and other Microsoft logos, as depicted in Fig. 9. All characters in the game are designed to resemble typical OS features and behavior: the main character, Tux, exhibits some power-ups (Linux enhancements) gained during the course of the game, usually as a reward for destroying enemies which, at their turn, mirror MS features. For instance, Office's paperclip *"Clippy"* behaves as an intrusive and annoying never-die item that bothers the penguin by constantly pursuing it. Destroying the clip (a very hard task because of its ability to show up and disappear randomly) gives the player extra points and other useful stuff. Although simple in conception, the game has proved to be extremely useful for students to recall the main features of some OS with no memorization at all. The videogame was implemented in licensed *Blitz3D* animation software on Linux. A Windows version is also available.

4 Conclusions and Further Remarks

This paper describes a new edutainment-oriented approach for powering students' engagement in the classroom at university and college levels. The approach relies on the intensive use of videogames and virtual reality technology in order to motivate students and make the courses amusing and engaging. The paper discusses some important issues related to our approach and show several examples of typical projects carried out during the course.

One of the typical reactions we found when explaining our edutainment approach was: *This is not serious. Such a learning-by-playing approach is fine for toddlers, not for university students.* In our opinion, there is a common misconception in the way videogames and virtual reality are regarded as educational tools. Very often they are considered as adequate *only* for children. The rationale behind is that they are too simple to be useful for middle or high-school students, not to mention those at college or university levels. Many people also believe that students at such levels (especially at university) neither need nor expect learning to be fun or engaging. But even if that statement would be true (and we do not really think so), most current students show a low level of motivation and engagement and get easily bored when difficult subjects such as maths or physics are taught. Our claim in this paper is that videogames and VR are valuable tools also for those students and subjects. It is simply a matter of which contents do you wish your students to learn and how these tools will be used in order to achieve your goals. This opinion is supported by our own experience as educators at university and our sudents' feedback, as evidenced in previous section.

This initiative is in some aspects similar to others recently reported in the literature. In fact, the educational videogame market has been expanding over the last five years [5]. We have recently seen a great deal of edutainment games aimed at all ages such as *Brain Training* or *Brain Age* by *Nintendo* and others [4]. Their commercial success clearly indicates that the approach is widely accepted and valuable for a number of subjects and people. In this paper we focused on university students firstly because we are university teachers, secondly because Bologna declaration is almost here and thirdly because of the few studies on edutainment performed at higher education level. But we really think this approach is rather ubiquitous and can be applied everywhere, including on-line learning and disabled people courses.

It is our hope that sharing our positive experience of teaching computer graphics by applying this edutainment approach will be of benefit to other educators and teachers. We would be delighted to receive the feedback and comments from others with similar experiences.

This research has been supported by the Computer Science National Program of the Spanish Ministry of Education and Science, Project Ref. #TIN2006-13615 and the University of Cantabria.

References

1. Allen, D.W., O'Shea, P.M., Baker, P.: Social and Cultural Foundations of American Education, 3rd edn (2007)
2. Di Blas, N., Poggi, C.: European virtual classrooms: building effective virtual educational experiences. Virtual Reality 11, 129–143 (2007)
3. The Bologna Declaration on the European space for higher education: an explanation. Association of European Universities & EU Rectors Conference, p. 4 (1999)
4. Cheung, K.K.F., Jong, M.S.Y., Lee, F.L., Lee, J.H.M., Luk, E.T.H., Shang, J., Wong, M.K.H.: FARMTASIA: an online game-based learning environment based on the VISOLE pedagogy. Virtual Reality 12, 17–25 (2008)
5. Egenfeldt-Nielsen, S.: Third generation educational use of computer games. Journal of Educational Multimedia and Hypermedia 16(3), 263 (2007)
6. Gálvez, A., Iglesias, A., Otero, C., Togores, R.: Matlab Toolbox for a First Computer Graphics Course for Engineers. In: Laganá, A., Gavrilova, M.L., Kumar, V., Mun, Y., Tan, C.J.K., Gervasi, O. (eds.) ICCSA 2004. LNCS, vol. 3044, pp. 641–650. Springer, Heidelberg (2004)
7. Gálvez, A., Iglesias, A.: Numerical-Symbolic Matlab Toolbox for Computer Graphics and Differential Geometry. In: Gervasi, O., Gavrilova, M.L., Kumar, V., Laganá, A., Lee, H.P., Mun, Y., Taniar, D., Tan, C.J.K. (eds.) ICCSA 2005. LNCS, vol. 3482, pp. 502–511. Springer, Heidelberg (2005)
8. Gálvez, A., Iglesias, A.: Matlab-Based Problem-Solving Environment for Geometric Processing of Surfaces. In: Iglesias, A., Takayama, N. (eds.) ICMS 2006. LNCS, vol. 4151, pp. 35–46. Springer, Heidelberg (2006)
9. Gálvez, A., Iglesias, A., Corcuera, P.: An Introductory Computer Graphics Course in the Context of the European Space of Higher Education: A Curricular Approach. In: Bubak, M., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2008, Part II. LNCS, vol. 5102, pp. 715–724. Springer, Heidelberg (2008)
10. Iglesias, A., Gutiérrez, F., Gálvez, A.: A Mathematica package for CAGD and Computer Graphics. In: Eurographics Workshop on Computer Graphics and Visualization Education, GVE 1999, Coimbra, Portugal, pp. 51–57 (1999)
11. Iglesias, A., Gálvez, A.: Effective BD-Binding Edutainment Approach for Powering Students Engagement at University Through Videogames and VR Technology. In: International Conference on Convergence Information Technology ICCIT 2008, Busan, Korea, pp. 307–314. IEEE Computer Society Press, Los Alamitos (2008)
12. Iglesias, A.: Facing the Challenges of the New European Space of Higher Education Through Effective Use of Computer Algebra Systems as an Educational Tool. RIMS Kokyuroku Journal Series, Special Issue on Computer Algebra Systems and Education: A Research about Effective Use of CAS in Mathematics Education 1624, 114–128 (2009)
13. Luengo, F., Contreras, M., Leal, A., Iglesias, A.: Interactive 3D Graphics Applications Embedded in Web Pages. In: International Conference on Computer Graphics, Imaging and Visualization-CGIV 2007, Bangkok, Thailand, pp. 434–440. IEEE Computer Society Press, Los Alamitos (2007)
14. Pan, Z., Chen, J.: Special Issue on VR-based Edutainment. Virtual Reality 12(1) (2008)

An SMS Spam Filtering System Using Support Vector Machine

Inwhee Joe* and Hyetaek Shim

Division of Computer Science and Engineering, Hanyang University,
Seoul, 133-791 South Korea
iwjoe@hanyang.ac.kr

Abstract. This paper describes a powerful and adaptive spam filtering system for SMS (Short Messaging Service) that uses SVM (Support Vector Machine) and a thesaurus. The system isolates words from sample data using a pre-processing device and integrates meanings of isolated words using a thesaurus, generates features of integrated words through chi-square statistics, and studies these features. The system is realized in a Windows environment and its performance is experimentally confirmed.

Keywords: Spam filtering system, short messaging service, support vector machine, thesaurus.

1 Introduction

Mobile phones are critical communications devices, and their associated SMS is used 1.5 to 2 times as much as voice service. As SMS usage increases, spam text messages are becoming more common. The average number of spam text messages received daily was reduced from 1.7 to 0.6 from December 2004 to May 2005, but increased to 0.74 in December 2005 and then to 0.99 in March 2006. Individuals classify mobile spam text messages as annoying (32.3%) time wasting (24.8%) and violating personal privacy (21.3%). Spam filtering functions installed on mobile phones identify specific number patterns or words and recognize spam messages when those numbers or words are present. However, this method cannot properly filter every type of spam message currently being dispatched. In this paper, we describe a novel protocol for structured content based spam filtering using SVM and a thesaurus, and explore multiple ways to optimize its performance. The background, significance, and structure of this paper are described in Section 1. In Section 2, we analyze traditional approaches and propose our novel spam filtering system. In Section 3, we describe the specifications and implementation of the spam filtering system. We implement the system in Section 4. In Section 5, we analyze our experimental results and discuss their implications.

* This work was supported by the ITRC Support Program (NIPA-2010-C1090-1021-0009) and the KEIT R&D Support Program of the MKE.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 577–584, 2010.

2 Related Work

Spam filtering is a peculiar filed to automatic document classification to considering the document is spam or not. Automatic document classification means make bunch of similar documents by allocate each document to proper category by get through the classification system.. That classification is consisting of two phases. First phase is feature selection method by extracting needed feature to classify after indexing bunch of documents. Second phase is decision make process that choose right category for the result from first phase.

Automatic document classification gets ability to assign right category automatically through mechanical learning process. For this process, it tagged specific word to bunch of learned document. The word represents the documents and extracting feature means batch job to select words revealed from learned document. However if it select every word in learned document as features, it takes too much time and looses judgment. To prevent this problem, calculate weight of information for each word then select featured words for automatic classification.

In text categorization, we are dealing with a huge feature spaces. This is why; we need a feature selection mechanism. The most popular feature selection methods are document frequency thresholding (DF) [1], the X 2 statistics (CHI) [2], term strength (TS) [3], information gain (IG) [1], and mutual information (MI) [1].

2.1 Chi-square Statistics

Chi-square statistics estimate the correlation between a specific word t and a category c and determine the difference between the observed value and the predicted value. A high chi-square value increases the chance that a feature will be selected [4].

$$X^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \tag{1}$$

a: number of documents containing word t among documents within category c b: number of documents containing word t among documents out of category c c: number of documents not containing word t among documents within category c d: number of documents not containing word t among documents out of category c.

This chi-square statistic has a natural value of zero if t and c are independent.

2.2 SVM(Support Vector Machine)

Standard support vector machines (SVMs) are powerful tools for data classification They classify two-category points by assigning them to one of two disjoint half spaces in either the original input space of the problem for linear classifiers, or in a higher dimensional feature space for nonlinear classifiers [5].

The role of a SVM is to construct a hyperplane as the decision surface such that the margin of separation between positive and negative examples is maximized. This desirable property is achieved by following a principled approach in statistical learning theory. More specifically, it uses a method of structural risk minimization. The theory uses the mathematical concept of Vapnik-Chervonenkis (VC) dimensionality and states that the generalization error rate is bounded by this term.

Optimal hyperplanes are constructed so that the VC dimension is minimized. The advantage of this technique is that good generalization performance is achieved for pattern classification problems without incorporating knowledge from the problem domain. This technique can be shown to correspond to a linear method in a high-dimensional feature space nonlinearly related to the input space [6].

Moreover, even though we can think of it as a linear algorithm in a high dimensional space, in practice, it does not involve any computations in that high dimensional space. Using kernels, all necessary computations are performed directly in the input space. The kernel function maps the input vector into a high dimensional dot product feature space implicitly and is used to construct the optimal hyperplane.

3 System Design

The proposed system composes three components.

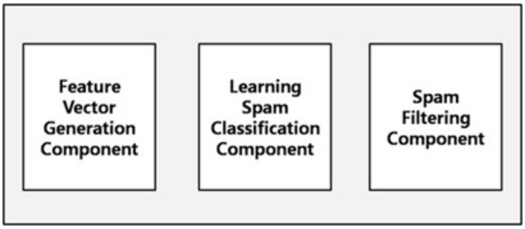

Feature Vector Generation Component
Learning Spam Classification Component
Spam Filtering Component

Fig. 1. Total Structure of Spam Filtering System

The first is a feature vector generator component that generates feature vectors after training. The second one is the SVM learner component using the generated feature vector. The last one is the spam filtering component to categorize spam messages using the completed classifier.

3.1 Feature Vector Generation Component

The feature vector extract component is a component that extracts feature vectors from learned data. The featured vector is a kind of array that marks 0 or 1 by the word existence. After extracting words through the preprocessor, we select the most heavily weighted word for judgment as a feature. The structure of the feature vector extraction component is as follows:

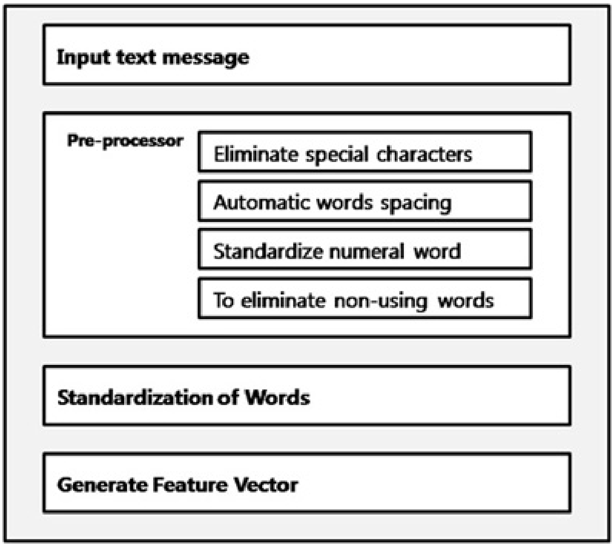


Fig. 2. Feature Vector Generation Component

The feature vector generator component goes through the following process. a. Pre-processing process b. Standardization of words using thesaurus c. Select features.

The pre-processing process then goes through the following 4 step process. - Eliminate special characters - Automatic words spacing - Standardize numeral word - Eliminate non-using words.

There are many cases in which spam SMSs include special characters. These special characters exist between words or characters, making the recognition of problem words impossible. Therefore, it is necessary to recognize words by eliminating special characters and spacing words automatically. The standardization of rhetoric and syllables recognizes both one thousand won and 1,000 won and uses it as a part of the feature vector. Non-used words are articles, prepositions, auxiliary words and conjunctions. They are eliminated since they are unnecessary. This pre-processing process is performed using the Korean language analysis module KLT [5]. The thesaurus is a word dictionary stored in a computer to search for information. It identifies special items showing synonyms, antonyms, and hyponyms. If there are synonyms among lists of words surveyed during pre-processing by the thesaurus, relevant words can be combined into one word and combined based on frequency for chi-square statistics. The feature vector was set to 100, 150, 200 and 300 in this study. The learning process was completed on each value. A higher chi-square statistic was selected as a feature and used as a learning component for the spam filtering system.

3.2 Learning Spam Classification Component

The learning component for spam classification generated learned vector data using character message data and putting it into the SVM classifier.

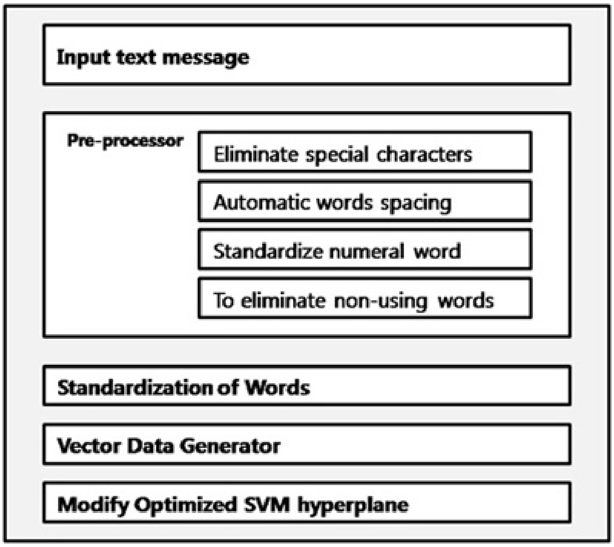


Fig. 3. Learning Spam Classification Component

The words of a text message are extracted while passing the preprocessor and standardized by the thesaurus. If the standardized word is in the feature list, the word index is set to 1 or 0. Generated vector values are used as learning data to modify the SVM hyperplane. Vector data are generated by SMS messages through two processes. If vector data has a matching word with an inserted SMS message, it marks 1 on the word. It then checks the stored contact address list to compare it with the contact address in the SMS message. If it finds a matching contact address, it marks 1 but 0. Lastly, by using information from standardizing the numeral word, if the SMS message contains money, it marks 1 or 0. After every feature vector is marked 0 or 1, a learning process is completed through SVM classifier. A Gaussian Radial Basis Function (RBF) is used as the kernel function. The constant value was set as 10, 20, 40, and gamma values were set at 0.01, 0.05, and 0.1 for this study.

3.3 Filtering Spam Component

The spam filtering component distinguishes whether the inserted data (SMS message) is spam or not by using the SVM classifier generated by the spam filter learning component. Words from the inserted SMS message are extracted by the preprocessor and standardized. Due to limitations of mobile devices, the

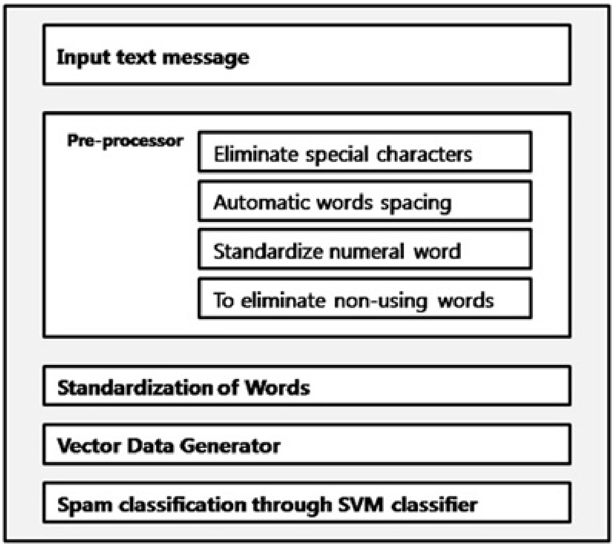


Fig. 4. Filtering Spam Component

Table 1. SPAM messages vs. Non-SPAM messages

	SPAM messages classified by system	Non-SPAM messages classified by system
Spam	a	c
Non Spam	b	d

thesaurus file extracts specific vector synonyms, and the digested thesaurus file is used for word standardization. Vector data is generated from extracted words, numeric information, and phone numbers. After putting the vector data into the SVM classifier, 0 indicates a non spam SMS message and 1 indicates spam.

4 Performance Evaluation

4.1 Performance Scaling Method

For performance scaling, this proposed system makes binary decisions in n ways following the chart below for document classification and information searching. (n = a + b + c + d)

$$\begin{aligned} \mathrm{SP} &= \frac{\text{Amount of SPAM message}}{\text{Total Amount of SPAM message}} \\ &= \frac{\mathrm{a}}{\mathrm{a+b}}(\text{if a+b} > 0) \end{aligned} \tag{2}$$

$$\mathrm{SR} = \frac{\text{Amount of real SPAM message}}{\text{Total Amount of SPAM message}} = \frac{\mathrm{a}}{\mathrm{a+c}}(\text{if a+c} > 0) \tag{3}$$

SP refers to Spam Precision, which is the ratio of correct to incorrect classifications among classified spam messages. SR refers to Spam Recall, which is the ratio of correct to incorrect predictions among real spam messages.

$$\mathrm{NSP} = \frac{\text{Amount of correct Non-SPAM message}}{\text{Total Amount of Non-SPAM message}} = \frac{\mathrm{d}}{\mathrm{c+d}}(\text{if c+d} > 0) \tag{4}$$

$$\mathrm{SR} = \frac{\text{Amount of correct Non-SPAM message}}{\text{Total Amount of Non-SPAM message}} = \frac{\mathrm{d}}{\mathrm{b+d}}(\text{if b+d} > 0) \tag{5}$$

NSP refers to Non-Spam Precision, which is the ratio of real to not real non-spam messages among those classified as non-spam messages. NSR refers to Non-Spam Recall, which is the ratio of non-spam messages classified correctly among whole real non-spam messages.

4.2 Experimental Results

Two hundred non-spam messages and 100 spam messages were used to train the system. Eighty spam messages and 80 non-spam messages, a total of 160 messages, were used for testing.

Table 2. Recognition Rate Per Number of Feature Vectors

Vector	SP	SR	NSP	NSR
100	95.89%	87.5%	88.5%	96.25%
150	93.58%	91.25%	91.46%	93.76%
200	93.24%	86.25%	87.2%	93.76%
300	86.66%	81.25%	82.35%	87.5%

Comprise recognition ratio during test with selected feature vector in higher order after apply chi-square statistic to the words. As shown in the chart, the most stable recognition result appears when the feature vector value is 150. Constant and gamma recognition tests were executed with feature vector value sets set to 150.

With a feature vector value of 150, the best level of recognition appears when the constant value is 20 and the gamma value is 0.01.

Table 3. Spam SMS Recognition Rate based on Gamma and Constant Values

Constant(Gamma)	SP	SR	NSP	NSR
10(0.01)	85.71%	75.02%	77.77%	87.5%
10(0.05)	89.33%	83.75%	84.70%	90.15%
10(0.1)	87.83%	81.25%	82.55%	88.75%
20(0.01)	91.56%	94.98%	94.80%	91.25%
20(0.05)	88.46%	86.25%	86.58%	88.75%
20(0.1)	91.78%	83.75%	85.05%	92.50%
40(0.01)	85.13%	78.75%	80.23%	86.25%
40(0.05)	87.67%	80.02%	81.60%	88.75%
40(0.1)	84.93%	77.50%	79.31%	86.25%

5 Conclusions

The spam filtering system proposed in this paper automatically sorts spam SMSs when an SMS is received based on the sender of the message and its content. The proposed system shows optimal performance with a feature vector value of 150, a constant value of 20 and a gamma value of 0.01 The proposed spam filtering system uses experience-based learning to recognize spam SMSs . Without training, it has a low recognition rate. This is a limitation of the machine learning algorithm that can be overcome by providing various patterns of learned data. The recognition rate was drastically reduced when the pre-processing device could not isolate word lines properly. Further study of automatic word spacing may be required.

References

1. Yang, Y., Pedersen, J.: A comparative study on feature selection in text categorization. In: The Fourteenth International Conference on Machine Learning (ICML 1997), pp. 412–420. Morgan Kaufmann, San Francisco (1997)
2. Schutze, H., Hull, D., Pedersen, J.: A comparison of classifiers and document representations for the routing problem. In: International ACM SIGIR conference on research and development in information retrieval (1995)
3. Yang, Y., Wilbur, J.: Using corpus statistics to remove redundant words in text categorization. Journal of the American Society of Information Science 47(5) (1996)
4. Greenwood, P.E., Nikulin, M.S.: A Guide to Chi-Square Testing. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken (2003)
5. Cortes, C., Vapnik, V.: Support vector network. Machine Learning 20, 273–297 (1995)
6. Sahay, S.: Support Vector Machines and Document Classification (2004)
7. `http://nlp.kookmin.ac.kr/HAM/kor/index.html`

Telecommunications User Behaviors Analysis Based on Fuzzy C-Means Clustering

Zhe Guo and Furong Wang

Electronic and Information Engineering Department of Huazhong University of Science and Technology, 430074 Wuhan, China
guozhe@foxmail.com, wangfurong@hust.edu.cn

Abstract. As the number of telecommunication user close to saturation, operators turn their focus from how to increase the subscriber to how to maintain the existing ones, which need more in-depth analysis of user character. In this paper we modeling user communication behavior based on the incoming/outgoing call holding time and then use fuzz c-means clustering algorithm to classify every level in user pyramidal model. For each level we get 3 classifications. We analyze the proportion and communication trend of each classification to help operators know their subscribers better. The method and conclusion of this paper can be used as the base of precision marketing for telecommunications industry.

Keywords: User Behavior, Data-Mining, Fuzz-C Means clustering.

1 Introduction

The telecommunications industry in China has been in a high-speed development from the beginning nineties of the last century, with China's economic boom. The telecommunications enterprise reform in that decade indicates that the government would like to see operators to compete, rather than monopoly. With the environment of increased competition in the market, telecom users are closing to saturation while in some metropolis the proportion of telecom-users even is more than 100% which means someone is served by more than one operator. These situation forces operators turn their attention from the development of new users to how to maintain the existing ones. To enhance the competitiveness, consumer behavior and customer relationship analysis must be more emphasized. Customer relationship management(CRM) is a broadly recognized, widely-implemented strategy for managing and nurturing a company’s interactions with clients and sales prospects. The overall goals are to find, attract, and win new clients, nurture and retain those the company already has, entice former clients back into the fold, and reduce the costs of marketing and client service[1]. The telecom operators want to distinguish between different users and in-depth understand the needs of different users groups in order to develop targeted marketing strategy that will improve the revenue by each type of user. The key of telecom CRM is customer segmentation which is the subdivision of a market into discrete customer groups that share similar characteristics, such as age, gender, interests, spending habits, and so on. Customer Segmentation can be a powerful means to identify unmet customer needs.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 585–591, 2010.

Currently, the customer segmentation method for telecom enterprise can be concluded as: demographic-based; behavior-based; value-based and attitude-based[2]. Actually, the most used method is based on ARPU(Average Revenue Per User). However, ARPU only reflects the value of a user and cannot distinguish between two users have similar ARPU. In this paper, we take combination of the value-based and behavior-based segmentation method to investigate the difference between users who have different or similar ARPU. That is, we divide user based on their ARPU first as the preliminary division, then consider the behavior—the minutes of outgoing/incoming call—of user based on fuzzy c-means algorithm(FCM). The result would be helpful to operator for market strategy.

This paper is organized as follows: Section 2 introduces user behavior indicator. In section 3, we provided a detail description of our proposed solution. The clustering result and analysis can be seen in section 4. Finally, section 5 concludes the paper.

2 Communication Behavior Indicator

Communication behavior indicator reflects the affordability, social role and calling habits of different user. Operators are more interested in behavior of the user's calling, thus referring to the concept of traffic we select the minutes of outgoing/incoming call per hour as user communication behavior indicator.

From h o'clock to h+1 o'clock, the minutes of outgoing call:

$$x_h = \sum_{i=1}^{n} t_{hi} , \tag{1}$$

where n is the number of outgoing call event, t_{hi} represents each call duration. Hence we get user j outgoing call vector of whole day:

$$X_j = \{x_1, x_2, \cdots, x_{24}\} . \tag{2}$$

Outgoing call vector provide a detailed statement about tendency of initiative communication request of user. Considering that many operators take one-way charge, outgoing call vector reflects the affordability of the user.

Similar to outgoing call vector, we define incoming call vector as:

$$Y_j = \{y_1, y_2, \cdots, y_{24}\} , \tag{3}$$

where y_i is the total minutes of incoming call at time i.

X_j and Y_j not only indicate the total traffic feature but also give details about the communication behavior. Figure.1 is a user's call duration about outgoing and incoming call of 24 hours a day (average of 3 months data, and normalized). Total duration of outgoing call is drawn in positive y axis while incoming negative y axis.

The pyramidal model of user's value is a common model used by operators. This model divides user by their ARPU, typically into 3 clusters which mean high, medium, low value respectively. However, even two users have similar ARPU, their behavior may be quite different—one would make call in the morning and the other prefer night. Outgoing/incoming call duration vector give us insight of this difference.

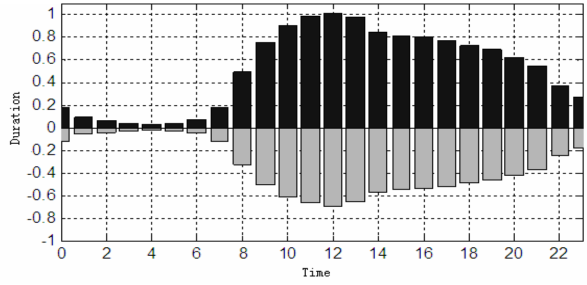


Fig. 1. Outgoing/incoming call duration of a day

3 Model of User Behavior

As mentioned in section 2, we get user outgoing/incoming call vector, X_j and Y_j. For n users, we get matrix X and Y composed as:

$$X= \{X_1,X_2,X_3,\dots X_n\}, \tag{4}$$

$$Y= \{Y_1,Y_2,Y_3,\dots Y_n\}. \tag{5}$$

X and Y are as the input of algorithm. Before data processing, we have to normalize the data in order to get a comparable result. In this paper, we use Min-Max normalization. Thus is, values are scaled such that the smallest value for each array becomes zero and the largest value becomes one.

The steps of our proposed method are as follows:

Step 1: Divide all sample users into three levels based on their ARPU (from high to low, the threshold is top20%, top50% and others);

Step 2: Normalize user data as input of FCM algorithm;

Step 3: Implement FCM algorithm for each level of user divided by ARPU, clustering into 3 classes for each level (c=3);

Step 4: Average each classes and calculate their proportion to corresponding level;

Here we give a brief introduce of FCM. FCM is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981[3-4]) is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^m \left\| x_i - c_j \right\|^2, \quad 1 \le \mathrm{m} \le \infty, \tag{6}$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center.

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_k\right\|}\right)^{\frac{2}{m-1}}}, \qquad c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m}. \tag{7}$$

This iteration will stop when $\max_{ij}\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas *k* are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m.

The algorithm is composed of the following steps:

Step 1. Initialize U=[u_{ij}] *matrix,* $U^{(0)}$

Step 2. At k-step: calculate the centers vectors $C^{(k)}$=[c_j] *with* $U^{(k)}$

Step 3. Update $U^{(k)}$, $U^{(k+1)}$

Step 4. If || $U^{(k+1)}$ - $U^{(k)}$||< ε *then STOP; otherwise return to step 2.*

The convergence of the algorithm has been proved[5].

The weighting exponent *m* is an important parameter in FCM algorithm. When *m* is close to one, the FCM approaches the hard c-means algorithm. When m approaches infinity, the only solution of the FCM will be the mass center of the data set. Hence, choosing a suitable weighting exponent is very important when implementing the FCM.

Bellman and Zadeh gave a optimization tool of choosing *m* [6]. Given a fuzzy objective function G and a constraint *C*, then a decision is produced by $D = G \cap C$. The membership function of data set *D* is defined as:

$$u_G(m) = \exp\{-\alpha \cdot \frac{J_m(U,P)}{\max_{\forall m}(J_m(U,P))}\}, \tag{8}$$

where α is a positive constant larger than 1, typically, α=1. 5.

The membership function of fuzzy constraint function *C* is:

$$u_C(m) = \frac{1}{1 + \beta \cdot (\frac{H_m(U,c)}{\max_{\forall m}(H_m(U,c))})}, \tag{9}$$

where β is a positive constant, typically, β=10.

Therefore, we get the optimized weighting exponent m^*:

$$m^* = \arg_{\forall m}\{\max\{\min\{u_G(m), u_C(m)\}\}\}. \tag{10}$$

As mentioned before, we get 3 level of user based on their ARPU. Implementing the algorithm of choosing optimized m^* is calculated as show in Figure.2:

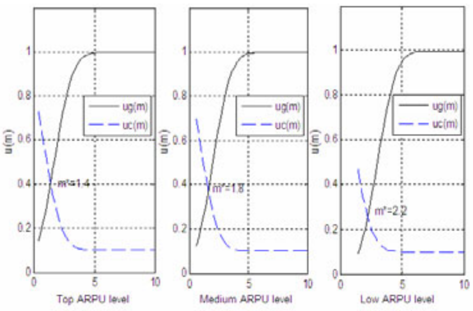


Fig. 2. Choose the optimized weighting exponent *m* for each level user

4 Clustering Result and Analysis

We get anonymized data which represents 3mothes of detailed bill from 10 thousands mobile phone users. The data contains the time and duration of every call generated or received. It also contains the total cost of each month.

Firstly, we divide users into 3 levels by their ARPU (from high to low). The proportion and total ARPU contribution of each level is show in Figure. 3.

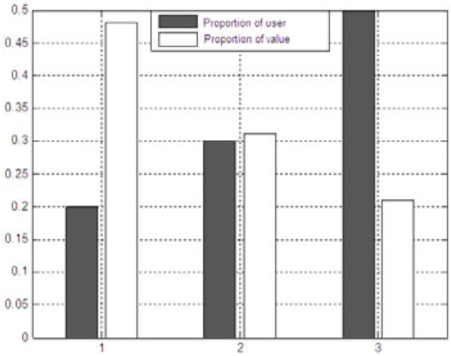


Fig. 3. The proportion and total ARPU contribution of each level

From Figure.3 we can see that due to the fierce competition, the gap of total contribution between each level is narrowed which indicates that we cannot underrate any level of user.

As mentioned in section 3, we get *c* and optimized m^*, then implement the FCM algorithm for each level of user, the clustering result is shown in Figure.4. The subgraph in Figire.4 presents different ARPU level, high ARPU to low. For convenience's sake, we define each cluster for a certain ARPU level as c1, c2 and c3 respectively. The proportion of c1, c2 and c3 is illustrated on the left of each subgraph.

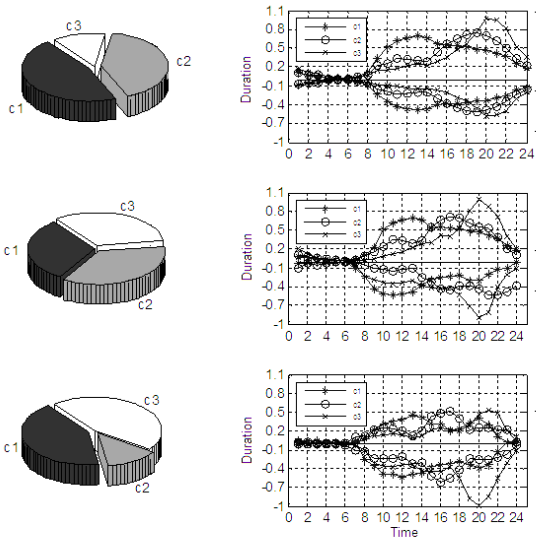


Fig. 4. Clustering result for each ARPU level

From Figure.4 we can see that for the top ARPU level, c1 has more call duration morning than afternoon or night, c2 less, c3 has a sharp peak in the night at 20:00. Generally speaking, the top ARPU level has more outgoing call minutes than incoming, which indicates they have strong affordability.

For the medium ARPU level, c1, c2, c3 has similar character as top ARPU level, besides, outgoing call minutes are approximately equal to incoming call.

For the low ARPU level, the characteristic of c1, c2, c3 is not as distinct as top or medium level—in FCM algorithm this level has the largest *m**—but based on incoming call minutes we can still reach a same conclusion as top and medium level. Besides, this level has more incoming call minutes obviously.

What's more, we have noted that c3 has a large ratio of population (about 45%) for medium and low ARPU level and about 40% of total population. This cluster makes the peak at night and would require attention of operators for it has great communication demand.

5 Conclusion

In this paper we have modeled telecom user behavior based on their outgoing/incoming call and implemented FCM algorithm after ARPU division. We get 3 clusters for each ARPU level, and find difference between each cluster. This difference can't be seen by ARPU division because they have similar ARPU. In addition, we are attracted to c3 cluster which has peak traffic in the night. In high ARPU level, the proportion of c3 is small but in medium and low ARPU is large. This kind of user would draw attention of operators because they are sensitive about price but have great communication demand.

References

1. Bhattacharya, C.B., Sen, S.: Consumer-Company Identification: A Framework for Understanding Consumers' Relationships with Companies. Journal of Marketing 67(2), 76–88 (2003)
2. Chris, R., Jyun-Cheng, W.C.Y.: David Data Mining Techniques for Customer Relationship Management. Technology in society 24(4), 493–502 (2002)
3. Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Cybernetics and Systems 3, 32–57 (1973)
4. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, Norwell (1981)
5. Bezdek, J.C.: A convergence theorem for the fuzzy ISODATA clustering algorithm. IEEE Transaction on Pattern Analysis and Machine Intelligence 1(2), 1–8 (1980)
6. Bellman, R.E., Zadeh, L.A.: Decision-Making in a Fuzzy Environment. Management Science 4(17), 141–164 (1970)

Performance Evaluation of a Hybrid TOA/AOA Based IR-UWB Positioning System

Nammoon Kim, Junho Gal, Eunyoung Lee, Inho Jeon, and Youngok Kim

Department of Electronics Engineering, Kwangwoon University, Korea
kimyoungok@kw.ac.kr

Abstract. In this paper, we evaluate performance of IR-UWB positioning system, which is based on the angle-of-arrival (AOA) and time-of-arrival (TOA) estimation techniques, with different IR-UWB waveforms, such as root raised cosine pulse, 5th order Gaussian mono-pulse, 4th modified Hermite pulse (MHP), and sine-type prolate shperoidal pulse. For ranging performance evaluation, the minimum mean square error (MMSE) technique is employed to resolve the multipath components, and the ranging performance is evaluated with various waveforms. For high precision angle estimation, the multiple signal identification and classification (MUSIC) method is employed. Simulation results show that the MHP pulse outperforms other considered waveforms in a hybrid TOA/AOA based IR-UWB positioning system.

Keywords: TOA, AOA, IR-UWB, MMSE, MUSIC.

1 Introduction

Recently, the high precision ranging techniques become an issue because of its various applications such as enhanced 911, U-health service, context aware service, navigation, high precision robot control and so on. In indoor environment, the time-of-arrival (TOA) and the angle-of-arrival (AOA) techniques are well known schemes for a high precision ranging system. Generally, the TOA based ranging scheme has better accuracy than the AOA based ranging scheme, because the AOA requires line-of-sight (LOS) wireless communication environment. However, the AOA system is more costly effective than the TOA system because the TOA system provides two-dimensional (2-D) positioning information with minimum three base stations (BSs), while the AOA system requires only two BSs. The proposed joint TOA and AOA positioning scheme requires only one BS for 2-D positioning information. By using the proposed hybrid TOA/AOA based IR-UWB positioning scheme, thus, we can precisely estimate the location of target with relatively less costs as compared with the existing methods using minimum three BSs or the estimating cell ID with one MS [1].

The TOA based ranging scheme experiences always multipath problem because of its signal's reflection, extinction, and so on. Once a signal passes multipath channels, a BS receives a combined signal that has many different phases [2]. To estimate more accurate distance, it is required to resolve the multipath channel accurately. In indoor positioning scheme, if ranging guarantees high precision, the performance of a hybrid

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 592–600, 2010.

TOA/AOA location scheme is highly trustful. Although the short pulse duration of IR-UWB waveform enables to resolve multipath channels accurately, the performance of ranging system under the condition of different waveforms is not studied enough. Since the IR-UWB waveform influences on the performance of IR-UWB ranging system, the ranging performance is evaluated with various waveforms in this paper. The MMSE are employed to resolve the multipath components for high precision ranging in the considered TOA scheme, while the MUSIC method is employed to measure the angle of target mobile terminal in the considered AOA scheme. The performance of proposed scheme is evaluated through the computer simulation over the channel models produced by IEEE 802.15.4a [3].

2 System Description

A simplified transceiver structure for TOA/AOA based IR-UWB positioning system is shown in Fig.1. The channel impulse response (CIR) can be estimated by applying the inversion or pseudo inversion of the known signal matrix. Then, the estimated channel matrix is applied to the TOA estimation process. Each antenna of antenna array receives an emitted signal with different delays caused distance gap on each antenna and the angle of target is estimated with the AOA scheme.

As mentioned in previous section, a pulse signal experiences multipath channel and a receiver receives an overlapped signal. Fig. 2 shows the comparison of various received signals according to interval of multipath in no noise communication condition. In the figure, we assumed that the pulse width is 2ns and the amplitudes of delayed signals are all the same. As shown in the figure, the multipath components are overlapped except for the Fig. 2-(d). The overlapped signals can cause to increase the error rates in ranging estimation.

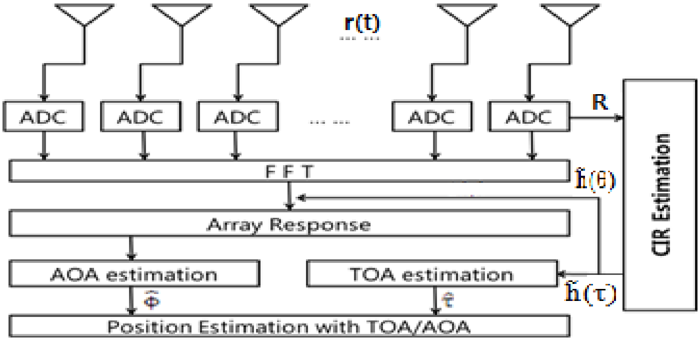


Fig. 1. The hybrid TOA/AOA positioning system design

Fig. 3 shows an example of received signal in a realistic TOA based IR-UWB ranging estimation system. As shown in the figure, the received signal consists of multiple multipath components and it is generally difficult to resolve the multipath components. If the overlapped multipath components can be resolved by using a received signal and already known transmitted signal, however, we can provide a higher precise TOA ranging performance.

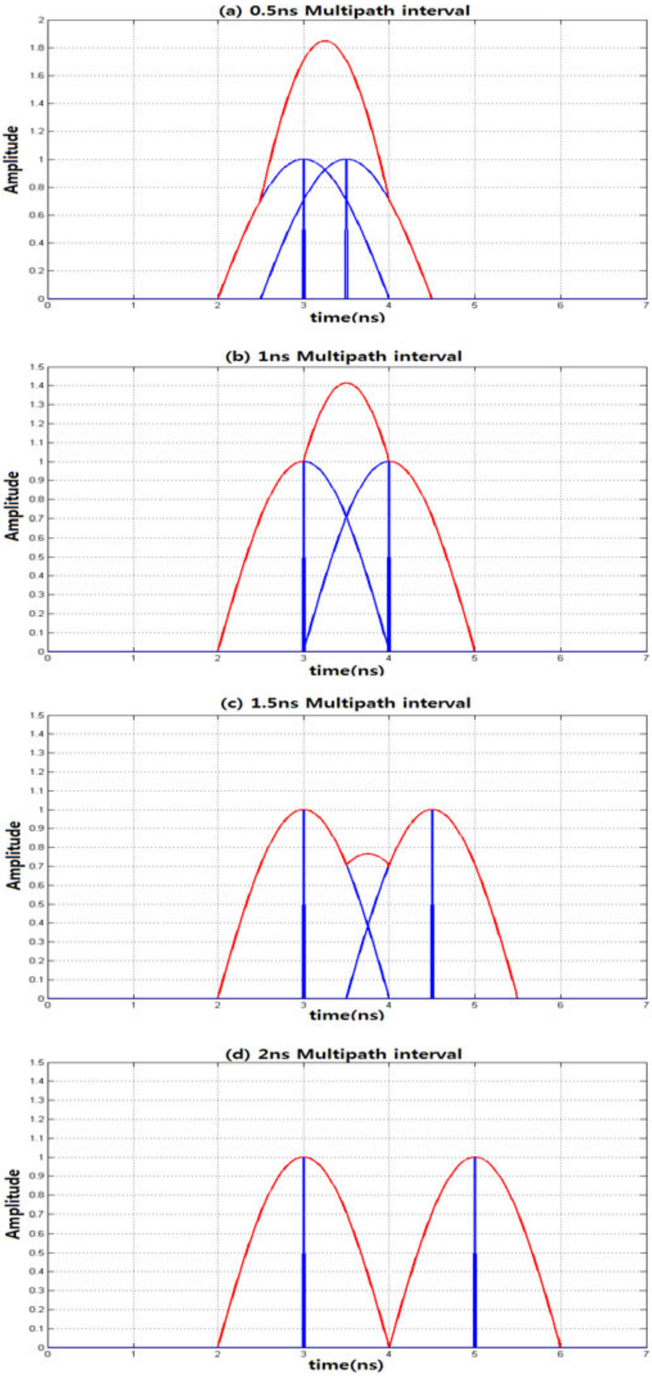


Fig. 2. Comparison of various received signals according to interval of multipath

At the receiver, the received signal over the multipath fading channel can be expressed as follows:

$$r(t) = \sum_{k=0}^{L_p-1} \alpha_k s(t-\tau_k) + w(t)\,, \tag{1}$$

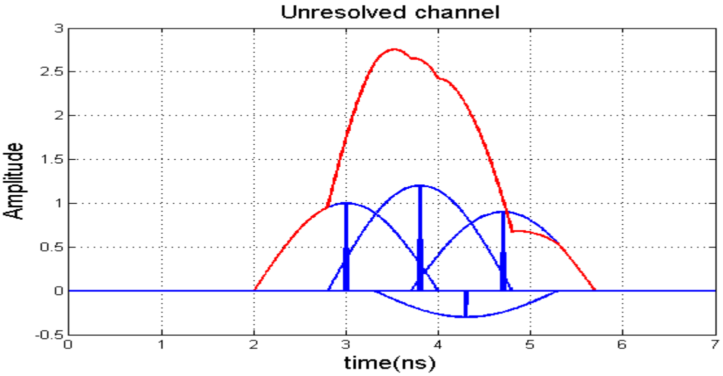


Fig. 3. Example of received signal with multiple multipath components

where L_p is the total number of multipath channels, α_k and τ_k are the amplitude and the propagation delay of the k-th path, respectively. s(·) is the transmitted pulse shape and w(t) is the additive white Gaussian noise with mean zero and variance σ_w^2 . By applying the harmonic signal model, (1) can be represented in frequency domain as follows [4][5]:

$$\begin{aligned} \mathrm{R(f)} &= \mathrm{S(f)H(f)} + \mathrm{W(f)} \\ &= \sum_{\mathrm{k=0}}^{\mathrm{Lp-1}} \alpha_{\mathrm{k}}\, \mathrm{S(f)} e^{-\mathrm{j}2\pi \mathrm{f}\tau_{\mathrm{k}}} + \mathrm{W(f)}\,, \end{aligned} \tag{2}$$

The discrete measurement data of (2) can be obtained by sampling at L equally spaced frequencies and is given by

$$R(f) = \sum_{k=0}^{L_p-1} \sum_{l=0}^{L\ \ -1} \alpha_k S(m+l) e^{-j2\pi(f_0+l\Delta f)\,\tau_k} + W(m)\,, \tag{3}$$

where m $= 0, 1, \cdots, M-1$, f_0 is center frequency, and Δf is the sampling interval in frequency domain. Since we use harmonic model in L frequency samples, the N samples are divided into M consecutive segments of length L, where $M = N - L + 1$. Therefore, the transmitted signal S is formed into a $M \times L$ matrix and the sampled signal of (3) can be rewritten as follows:

$$\mathbf{R} = \mathbf{SH} + \mathbf{W} = \mathbf{SVa} + \mathbf{W}, \tag{4}$$

where

$$\mathbf{R} = [R(0) \quad R(1) \quad \cdots \quad R(M-1)]^{\mathrm{T}},$$

$$\mathbf{S} = \begin{bmatrix} S(0) & S(1) & \dots & S(L-1) \\ S(1) & S(2) & \cdots & S(L) \\ \vdots & \vdots & \ddots & \vdots \\ S(M-1) & S(M) & \cdots & S(M+L-2) \end{bmatrix},$$

$$\mathbf{H} = [H(f_0) \quad H(f_1) \quad \cdots \quad H(f_{L-1})]^{\mathrm{T}},$$

$$\mathbf{W} = [W(0) \quad W(1) \quad \cdots \quad W(M-1)]^{\mathrm{T}},$$

$$\mathbf{V} = \left[\mathbf{v}(\tau_0) \quad \mathbf{v}(\tau_1) \quad \cdots \quad \mathbf{v}(\tau_{L_p-1})\right],$$

$$\mathbf{v}(\tau_k) = [1 \quad e^{-j2\pi\Delta f\tau_k} \quad \dots \quad e^{-j2\pi(L-1)\Delta f\tau_k}]^{\mathrm{T}},$$

$$\mathbf{a} = \left[\alpha_0 e^{-j2\pi f_0\tau_0} \quad \alpha_1 e^{-j2\pi f_0\tau_1} \quad \cdots \quad \alpha_{L_p-1} e^{-j2\pi f_0\tau_{L_p-1}}\right]^{\mathrm{T}}.$$

3 Channel Models and IR-UWB Pulses

3.1 Channel Models

The CM1 and CM3 of IEEE 802.15.4a UWB channel model are considered. The CM1 represents a LOS environment of less than 4m and the CM3 is for a LOS environment of 4-10m. In this paper, we assumed the direct path is the multipath component having the strongest amplitude because of LOS environment.

3.2 Various IR-UWB Pulses

Various IR-UWB pulses are considered for IR-UWB positioning system. We employed the RRC pulse, the 5th order Gaussian mono-pulse, the 4th MHP pulse [6], and a sine type of PS pulse [7].

4 Channel Impulse Response Estimation

The CIR can be estimated by applying the inversion or pseudo inversion of the known signal matrix. By multiplying both sides of (4) by the inverse of the signal shape matrix $\mathbf{S}^{+}$, where $\mathbf{S}^{+} = \mathbf{S}^{\mathrm{H}}\{\mathbf{S}\cdot\mathbf{S}^{\mathrm{H}} + (\sigma_{\mathrm{w}}^{2})\cdot\mathbf{I}\}^{-1}$is for MMSE, it can be rewritten as follows:

$$\mathbf{S}^{+}\mathbf{R} = \mathbf{S}^{+}\mathbf{SH} + \mathbf{S}^{+}\mathbf{W} \ \text{ or } \ \widetilde{\mathbf{H}} = \mathbf{H} + \widetilde{\mathbf{W}}, \tag{5}$$

where I represents an identity matrix. Then, the estimated channel matrix $\widetilde{\mathbf{H}}$ is applied to the TOA estimation process and the frequency response of the estimated noisy CIR from (5) can be written as follows [4]:

$$H(j2\pi l \bigtriangleup f) = \sum_{k=0}^{L_p-1} \alpha_k z_k^l + W_k \ , \tag{6}$$

where $z_k = e^{-j2\pi\bigtriangleup f\tau_k}$ with $\bigtriangleup f = 1/L \cdot\bigtriangleup t$.

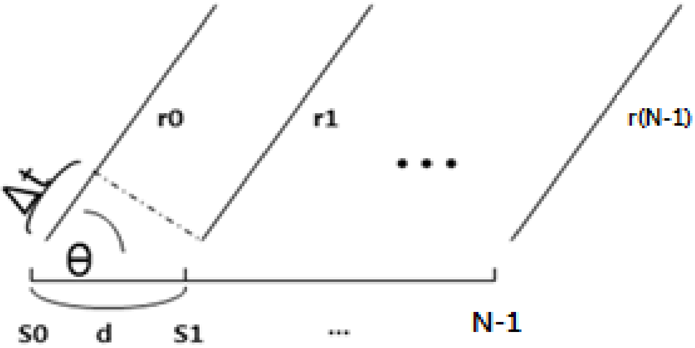


Fig. 4. Difference of detected signal depending on location of antenna in array antennas

5 AOA Estimation Using MUSIC

The direction of propagation of a radio-frequency wave can be estimated by the antenna array. Fig. 4 indicates the difference of detected signals depending on location of an antenna in array antennas. When a transmitted signal arrives to array antennas, each antenna receives the signal with different delays caused from distance gap on each antenna. The received signals between antennas are phase-shifted and they can be defined as follows:

$$\mathrm{e}^{-\mathrm{j}2\pi\mathrm{f}\Delta\mathrm{t}} = \mathrm{e}^{-\mathrm{j}2\pi\mathrm{fdcos}\theta/\mathrm{c}} = \mathrm{e}^{-\mathrm{j}2\pi\mathrm{dcos}\theta/\lambda} = \mathrm{e}^{-\mathrm{jkdcos}\theta} \left(\mathrm{k} = \frac{2\pi}{\lambda}\right), \tag{7}$$

and the relation of detected signals on S0 and S1 is expressed as follows; $\mathrm{S1} = \mathrm{S0e}^{-\mathrm{jkdcos}\theta}$. Thus, the detected signals on the array antenna can be defined as steering vector given by [8]:

$$\mathrm{s}(\theta) = \left[1 \quad \mathrm{e}^{-\mathrm{jkdcos}\theta} \quad \mathrm{e}^{-2\mathrm{jkdcos}\theta} \quad \cdots \quad \mathrm{e}^{-\mathrm{j(N-1)kdcos}\theta}\right]^{\mathrm{T}}, \tag{8}$$

where N is number of antennas in array.

5.1 MUSIC Algorithm

The MUSIC super-resolution techniques are based on eigen-decomposition of the autocorrelation matrix of the received signal vector. The received signal vector can be expressed as follows:

$$\mathrm{R} = \mathrm{SH} + \mathrm{W}, \tag{9}$$

where

$$S= [s(\theta_1)\, s(\theta_2)\ \cdots\ s(\theta_M)],$$
$$H = [\mathrm{H}_1\mathrm{H}_2 \cdots\ \mathrm{H}_M]^T,$$

and M is number of multipaths. The matrix S is N × M matrix.
The autocorrelation matrix of received signal is expressed as follows:

$$\mathrm{R}_{RR} = \mathrm{E}\{\mathrm{R}R^H\} = \mathrm{VA}V^H + \sigma_w^2\mathrm{I} = \mathrm{R}_s + \sigma_w^2\mathrm{I}, \tag{10}$$

where

$$\mathrm{R}_s = \mathrm{VA}V^H$$

$$\mathrm{A} = \begin{bmatrix} E[|H_1|^2] & 0 & \cdots & 0 \\ 0 & E[|H_2|^2] & \cdots & 0 \\ 0 & 0 & \cdots & E[|H_M|^2] \end{bmatrix}$$

The signal covariance matrix, R_s, is clearly a N × N matrix with rank M. Therefore, it has N-M eigenvectors corresponding to the zero eigenvalues. Let q_m be such an eigenvector. Then,

$$\begin{aligned} \mathrm{R}_s\, q_m &= \mathrm{SA}S^H q_m = 0, \\ &\Rightarrow \mathrm{q}_m^H \mathrm{SA}S^H q_m = 0, \\ &\Rightarrow S^H q_m = 0 \end{aligned} \tag{11}$$

The last equation of (11) is valid since the matrix A is clearly positive definite. The equation (11) implies that all the N – M eigenvectors (q_m) of R_s corresponding to the zero eigenvalues are orthogonal to all the M signal steering vectors. Let $Q_n$is the $\mathrm{N} \times (\mathrm{N}-\mathrm{M})$ matrix of these eigenvectors, then the MUSIC plots the pseudo-spectrum as follows:

$$P_{MUSIC(\theta)} = \frac{1}{\sum_{m=1}^{N-M}|S^H(\theta)q_m|^2} = \frac{1}{\mathrm{s}^H(\theta)Q_n\, Q_n^H s(\theta)} = \frac{1}{\left|Q_n^H s(\theta)\right|^2}, \tag{12}$$

Since the eigenvectors making up Q_n are orthogonal to the signal steering vectors, the denominator becomes zero when θ is a signal direction. Therefore, the estimated signal directions are the M largest peaks in the pseudo-spectrum.

6 Simulation Results

In this section, we evaluated the performance of proposed hybrid TOA/AOA based IR-UWB positioning system with various waveforms. As channel models, the CM1 and the CM3 of IEEE 802.15.4a standard were employed for computer simulations.

Fig. 5 depicts the position estimation error of hybrid TOA/AOA based IR-UWB system. As shown in the figures, the positioning system with MHP pulse outperforms the system with all other pulses for both CM1 and CM3 in high SNR region. Note that the positioning performance with other pulses is not remarkably enhanced as the SNR increase in CM3 while the positioning performance is enhanced as the SNR increase in CM1.

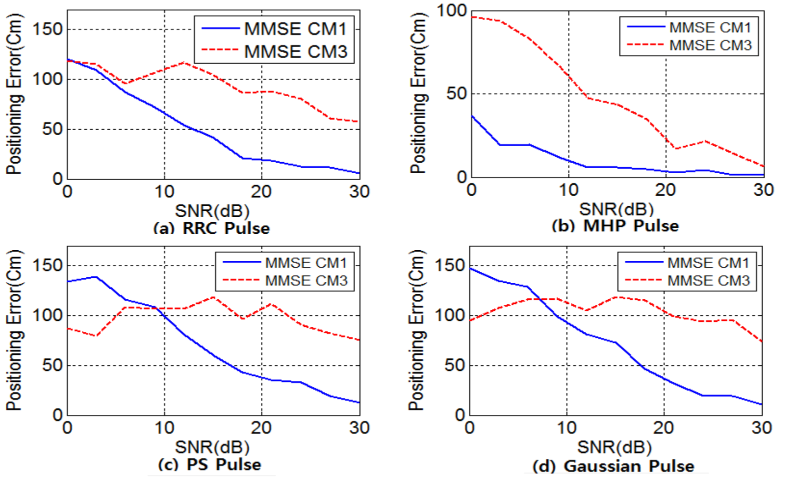


Fig. 5. Positioning estimation error of hybrid scheme with various waveforms

7 Conclusion

In this paper, we evaluated performance of hybrid TOA/AOA based IR-UWB positioning system with various shapes of waveforms. In a hybrid TOA/AOA scheme, it is shown that the MHP pulse outperforms all the pulses for both CM1 and CM3 in high SNR region. Although the accuracy of TOA scheme is better than that of hybrid TOA/AOA scheme, the hybrid TOA/AOA based IR-UWB positioning system requires only one BS, unlike the TOA system does three BSs. If the required accuracy of positioning system is tens of centimeters, the hybrid TOA/AOA based IR-UWB positioning system with MHP pulse can be considered because of its high performance and economic effects.

Acknowledgment

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.2010-0008494).

References

[1] Kim, N.Y.: Research on Computationally Efficient Network-based Wireless Geolocation Systems, doctoral dissertation, Dept. of Information and Communication Engineering, Korea Advanced institute of Science and Technology (2010)

[2] Taylor, J.D.: Introduction to Ultra-Wideband Radar Systems. CRC Press, Boca Raton (1995)

[3] IEEE 802.15 WPAN Low Rate Alternative PHY Task Group 4a, PART 15.4:Wireless MAC and PHY Specifications for LR-WPANs, Draft P802.15.4a/D7 (March 2007)

[4] Manolakis, D., Ingle, V., Kogon, S.: Statistical and Adaptive Signal Processing. McGraw-Hill, New York (2000)

[5] Kim, N.Y., Kim, S., Kim, Y., Kang, J.: A High Precision Ranging Scheme for IEEE802.15.4a Chirp Spread Spectrum System. IEICE Transactions on Communications E92-B(3), 1057–1061 (2009)

[6] Michael, L.B., Ghavami, M., Kohno, R.: Multiple pulse generator for ultra-wideband communication using Hermite polynomial based orthogonal pulses. In: Proc. IEEE Conf. UltraWideband Syst. Technol., pp. 47–51 (May 2002)

[7] Yin, L., Hongbo, Z.: Interference Mitigation in UWB Communications through Pulse Waveform Design. In: Environmental Electromagnetics, the 2006 4th Asia-Pacific Conference, August 1-4, pp. 569–572 (2006)

[8] Gross, F.B.: Smart Antennas for Wireless Communications with Matlab. McGraw-Hill, New York (2005)

All-Optical RZ-to-NRZ Converted Data Transmission at 10 Gb/s

Hyuek Jae Lee[1], Ik Soo Jin[1], and Hae Geun Kim[2]

[1] Dept. Of Information & Communication Engineering, Kyungnam University,
449 Wolryeong-dong, Masanhappo-gu, Changwon-si, Korea
[2] School of Computer and Information Communication, Catholic University of Daegu,
330 Kumrak-ri, Hayang-up, Kyungsan-si 712-702, Korea
{Hyuek,isjin}@Kyungnam.ac.kr

Abstract. This paper proposes and demonstrates a novel all-optical return-to-zero (RZ) to nonreturn-to-zero (NRZ) data format conversion using a semiconductor optical amplifier (SOA) loop mirror. The format conversion has been performed between the most widely used data formats—NRZ and RZ formats. The format conversion scheme is based on gain variation by an intensity-dependent phase change in an SOA-loop mirror. 10 Gb/s error-free fiber transmission up to 78 km for the converted NRZ format data is achieved. Further, the proposed method shows improved transmission performance than the conventional Mach-Zehnder modulation technique.

Keywords: All-optical RZ-to-NRZ, SOA(semiconductor optical amplifier), optical loop mirror, data format conversion.

1 Introduction

Future all-optical networks are likely to employ both wavelength division multiplexing (WDM) and optical time division multiplexing (OTDM) and there will be a need for all-optical data format conversion between WDM and OTDM signals [l]. Non-return-to-zero (NRZ) and return-to-zero (RZ) formats are both widely used data formats. While the RZ format is preferred in ultra-fast OTDM networks to make use of bit-interleaving technique, the NRZ format has a lower bandwidth requirement and a higher timing jitter tolerance than RZ format. Therefore, RZ-to-NRZ format converter is essential in linking and interfacing the ultra-fast OTDM networks and the lower speed WDM networks [1]. Previous reports included all-optical RZ-to-NRZ conversion using a Mach-Zehnder PIC [2], SOA/fiber grating filter [3], SOA-XGM [l], and NOLM [4]. None except for Ref. [2] reported optical fiber transmission of the converted NRZ data. Also we propose and demonstrate, for the first time to the best of our knowledge, 10 Gb/s optical fiber transmission of the RZ-to-NRZ converted data using SOA-loop mirror.

2 RZ-to-NRZ Conversion Using SOA-Loop-Mirror

Fig. 1 shows an experimental setup for the proposed RZ-to-NRZ converter scheme. A nonlinear optical loop mirror using an SOA (SOA- loop-mirror) is often used for

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 601–605, 2010.

all-optical switching in OTDM networks [5]. The switching principle is based on the optically induced phase difference (that is controlled by external control pulses) between a clockwise (*cw*) and a counter-clockwise (*ccw*) pulse in a fiber loop mirror.

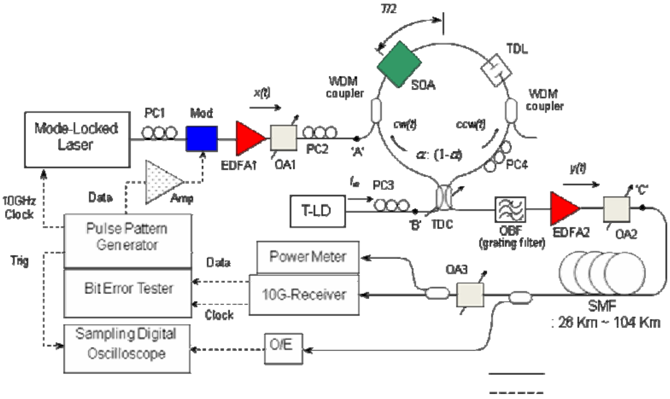


Fig. 1. Experimental Setup

Due to the displacement, it executes an exclusive OR (XOR) function for the input RZ signal and its *T* delayed signal, and finally generates NRZ signal. From the SOA-loop-mirror of Fig.1, the transmitted intensity is given by [6]

$$y = I_{in}(a^2 G_{cw} + (1-a)^2 G_{ccw} - 2a(1-a)\sqrt{G_{cw}G_{ccw}}\cos\theta_{diff}) \tag{1}$$

where, I_{in} is the input intensity, a is the coupling coefficient of the TDC, G_{cw} and G_{ccw} are the gains of an SOA for *cw* and *ccw* beams, respectively, and $\theta_{diff}\,(=\theta_{cw} - \theta_{ccw})$ is the phase difference between *cw* and *ccw* beams. For the simplicity, we set a TDC coupling coefficient a to 0.5. The operational timing diagram is illustrated in Fig. 2. In fact, the gain and the phase of *cw* and *ccw* beams depend on the intensity of the input signal *x(t)*, simultaneously. Here, we assume that when the input signal *x(t)* is 'on', the gain of the *cw* beam is the same as that of the *ccw* beam due to cross-gain modulation (XGM) of the SOA, i.e., $G_{cw,on} = G_{ccw,on}$, and when the input is 'off', $G_{cw,off} = G_{ccw,off}$. Cross-phase modulation (XPM) of the SOA also induces the phase difference between $I_{in}G_{cw,on}$ (or $I_{in}G_{cw,off}$) and $I_{in}G_{ccw,off}$ (or $I_{in}G_{ccw,on}$), which is assumed to be 'π' (or '-π') as shown in Fig. 2.

If we consider only the XPM effect (the gain is constant (=*G*)) alone, the output of the SOA-loop-mirror can be expressed as

$$\begin{aligned} y(t) &= 0.5 I_{in} G(1-\cos(\theta_{cw}(t) - \theta_{ccw}(t))) \\ &= 0.5 I_{in} G(1-\cos(\theta_{cw}(t) - \theta_{cw}(t+T))) \end{aligned} \tag{2}$$

On the contrary, with only the XGM effect (that is, the phase is constant ($\theta_{cw} = \theta_{ccw} = \theta$)), Eq. (1) is reduced to

$$\begin{aligned} y(t) &= 0.5 I_{in}(G_{cw}(t)+G_{ccw}(t)-2\sqrt{G_{cw}(t)G_{ccw}(t)}) \\ &= 0.5 I_{in}(G_{cw}(t)+G_{cw}(t+T)-2\sqrt{G_{cw}(t)G_{cw}(t+T)}) \end{aligned} \tag{3}$$

From (2) and (3), if $\theta_{cw}(t)= \theta_{ccw}(t)$ $(\theta_{cw}(t+T))$ and/or $G_{cw}(t)=G_{ccw}(t)$ $(=G_{cw}(t+T))$, $y(t)=0$. Otherwise, $y(t)$ has a certain value. Therefore, we can obtain the relation of $y(t) = x(t) \oplus x(t+T)$, where $\oplus$ denotes XOR logic. In fact, because the phase and/or the gain is not able to be abruptly changed, the output $y(t)$ shows a NRZ data format rather than an RZ data format like the doted output in Fig. 2.

The operation principle of the proposed RZ-to-NRZ conversion is topologically identical to that of Ref. [2] except for the use of a grating filter to obtain good chirping characteristics as shown in fig. 1. Ref. [2] employs a scheme using a Mach-Zehnder (MZ) interferometer while the proposed scheme is based on a Sagnac interferometer. In Fig. 1, the incoming RZ signal $x(t)$ enters into the fiber loop mirror through the WDM coupler. Because the SOA is located at the displacement of $\tau/2$ ($\tau\leq T$, T is a signal period) from the mid point of the fiber loop, the *ccw* beam takes a phase changing effect τ later than the *cw* beam from the incoming RZ signal as shown in Fig. **2.** Due to the phase change difference in both *cw* beam and the τ-delayed *ccw* beam, the τ gating window can be made. This operational principle is topologically identical to Ref. [2], but the proposed scheme has more simple and stable architecture.

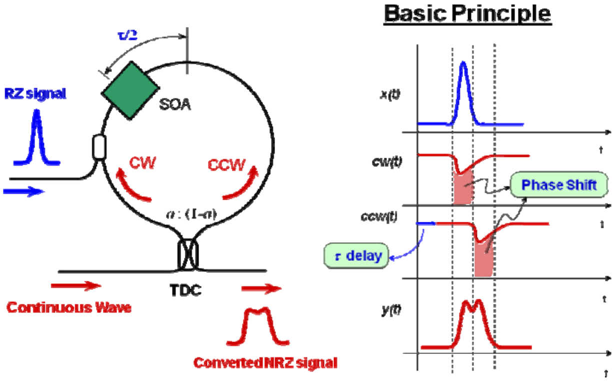


Fig. 2. Principle of the proposed RZ-to-NRZ conversion

3 Experiments and Results

In the experimental setup of Fig. 1, the SOA-loop-mirror consists of TDC, TDL, PC4, and SOA, which are connected by the WDM coupler for the RZ input signal $x(t)$. A $LiNbO_3$ electro-optic modulator and a mode-locked laser (~6 psec FWHM, 10 GHz) driven by a pulse pattern generator (PPG), generates a 10 Gb/s 2^{31}-1 RZ data sequence at 1557.13 nm. The RZ input signal enters into the SOA-loop-mirror through the WDM coupler. For NRZ data generation, the continuous wave beam I_{th} at 1550 nm is generated from Tunable-LD. The SOA used in this experiment was 1000 μm-long and nearly polarization insensitive (~0.6 dB). The SOA current is set to 190 mA, coupling coefficient α of TDC was adjusted to 0.41. The SOA-arrival time difference τ was set to ~70 psec. The optical power at points 'A', 'B' and 'C' in Fig. 1 was set to 8.7 dBm, 5.0 dBm, and 8.5 dBm , respectively.

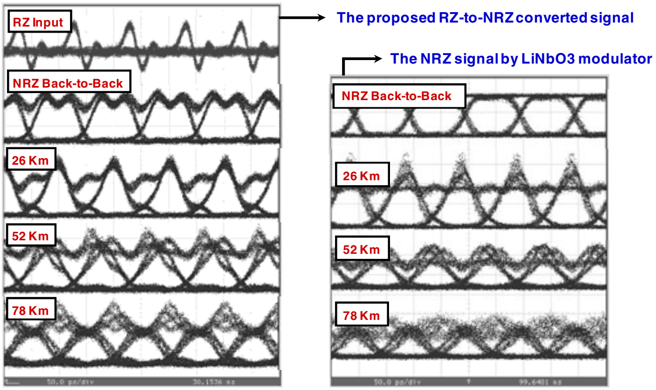


Fig. 3. Principle Eye diagrams of the optical transmission for of the proposed RZ-to-NRZ conversion (a) the proposed RZ-to-NRZ converted signal and (b) the conventional NRZ signal generated by LiNbO$_3$ MZ-modulator

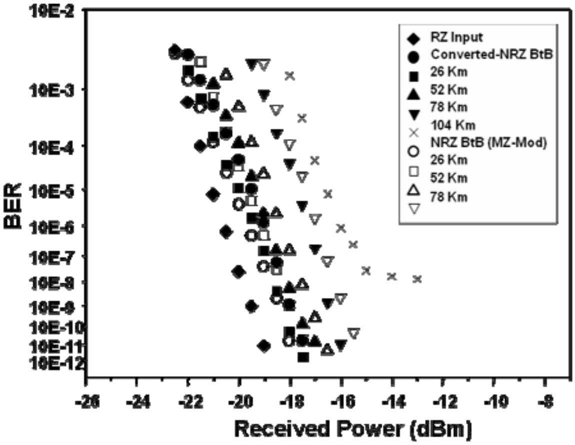


Fig. 4. Comparison of the measured BERs for the proposed RZ-to-NRZ signal and the conventional NRZ signal according to fiber transmission lengths

Fig. 3(a) shows eye diagrams for 10Gb/s RZ-to-NRZ converted signal and its optical transmissions over 26 km to 78 km dispersive standard single mode fiber. For comparison, the eye diagrams for the conventional NRZ signal generated by LiNbO$_3$ MZ-modulator are also shown in Fig. 3(b). Note that in Fig. 3, the propagation eyes for the proposed method are better than those for the conventional NRZ modulation method even though the original RZ-to-NRZ converted signal is distorted. BERs for

26, 52, 78, and 104 km fiber transmissions are shown in Fig. 4 and the proposed RZ-to-NRZ converted signal show improved transmission performances compared to the conventional NRZ signal.

4 Summary

We have proposed and demonstrated a novel all-optical RZ-to-NRZ data format conversion using SOA-loop-mirror. 10 Gb/s error-free transmission up to 78 km for the converted NRZ format data is achieved. Also, the proposed RZ-to-NRZ conversion method shows improved fiber transmission performance than the conventional Mach-Zehnder modulation method, and will serve as a key block in linking and interfacing for all-optical OTDM and WDM networks.

Acknowledgments. This work was supported by Kyungnam University Foundation Grant, 2010.

References

1. Norte, D., Park, E., Willner, A.E.: All-optical TDM-to-WDM data format conversion in a dynamically reconfigurable WDM network. IEEE Photon. Technol. Lett. 7, 920–922 (1995)
2. Park, S.G., Spiekman, L.H., Eiselt, M., Wiesenfeld, J.M.: Chirp consequence of all-optical RZ to *NRZ* conversion using cross-phase modulation in an active semiconductor photonic integrated circuit. IEEE Photon. Technol. Lett. 12, 233–235 (2000)
3. Cho, P.S., Mahgerefteh, D., Goldhar, J.: 10Gbh RZ to NRZ format conversionusing a semiconductor-optical-amplifier/fiber-bragg-grating wavelength converter. In: ECOC 1998, pp. 353–354 (1998)
4. Bigo, S., Desurvire, E., Desruelle, B.: All-optical RZ-to-NRZ format conversion at l0 Gbit/s with nonlinear optical loop mirror. Electronics Lett. 30, 1868–1869 (1994)
5. Lee, H.J., Kim, H.G., Choi, J.Y., Kim, K., Lee, J.: A simple packet-level clock extraction scheme using a terahertz optical asymmetric demultiplexer. IEEE Photon. Technol. Lett. 11, 1310–1312 (1999)
6. Eiselt, M., Pieper, W., Weber, H.G.: SLALOM: Semiconductor laser amplifier in a loop mirror. J. Lightwave Technol. 13, 2099–2112 (1995)

Improved Location Aided Cluster Based Routing Protocol for GPS Enabled MANETs

S. Mangai[1] and A. Tamilarasi[2]

[1] Department of Electronics & Communication Engineering,
Velalar College of Engineering and Technology,
Thindal, Erode, Tamilnadu-638 012, India
[2] Department of Computer Science and Engineering,
Kongu Engineering College,
Perundurai, Erode, Tamilnadu-638052, India
ishamangai@yahoo.com

Abstract. Routing has been the main challenge for ad hoc networks due to dynamic topology as well as resource constraints. Completely GPS free as well as GPS scarce positioning systems for wireless, mobile, ad-hoc networks has been proposed recently by many authors. High computational overhead and high mobility of the nodes typically require completely GPS enabled MANETs for higher performance. In this paper, Improved Location aided Cluster based Routing Protocol (ILCRP) for GPS enabled MANETs has been evaluated for performance metrics such as end to end delay, control overhead, and packet delivery ratio. Use of cluster based routing as well as exact location information of the nodes in ILCRP reduces the control overhead resulting in higher packet delivery ratio. GPS utility in nodes reduces the end to end delay even during its high mobility. Simulations are performed using NS2 varying the mobility (speed) of nodes as well as number of the nodes. Results illustrate that the ILCRP performs better compared to other protocols.

Keywords: GPS, NS2, Location aided routing, Cluster based routing.

1 Introduction

“Resource Constraint” is an extreme challenge faced by a routing protocol designed for ad hoc wireless networks. Gadgets used in the ad hoc wireless networks in most cases require portability and hence they also have size and weight constraints along with the restrictions on the power source. Control overhead increases due to mobility of the nodes resulting in bandwidth constraint. Mobility also affects end to end delay as well as packet delivery ratio. Therefore in real time applications there is a reduction in quality due to Bandwidth constraint. As a result, ad hoc network routing protocol must optimally balance these contradictory aspects.

Many routing protocols [1] have been proposed to reduce the complexity of a flat structured routing either with help of the clustering schemes or using location information of the nodes.A cluster based structure, synonymous to hierarchical routing, is obtained by Virtual partitioning of MANETs, each containing a cluster head node, a gateway node with inter cluster links and member nodes.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 606–615, 2010.

2 Related Work

Many algorithms have been proposed to optimize the procedure for election of cluster head. Lowest-ID algorithm (LID) [2-3] uses minimum ID whereas Highest-Degree (HD) [4] uses degree of the node as a metric for cluster head election. The degree of a node is the number of neighbor nodes. Since LID chooses the lowest ID node as the cluster head, it drains the cluster head's battery at a faster rate which perturbs the cluster stability. In HD algorithm a node having highest degree of connectivity is selected as the cluster head. Although HD algorithm reduces the delay as well as the number of clusters, it increases reaffliation overhead resulting in reduced throughput.

Mobility Metric Based Algorithm(MOBIC) [5], a variation of Lowest-ID algorithm yields better cluster stability at the cost of higher delay by choosing a relatively low mobile node as cluster head.

Node mobility as well as transmission range are taken for weight calculation in Distributed Mobility Adaptive Algorithm(DMAC) [6].Most of the algorithm such as Weighted Clustering Algorithm(WCA) [7-9], Generalized Distributed Mobility Adaptive Clustering(GDMAC) [10] are derived from DMAC.WCA considers degree of connectivity, mobility, battery power and transmission power. WCA is extended to improve performances in IWCA[11], FWCA[12].GDMAC improves the performance by introducing a cluster density parameter for the whole network.FWCA focuses on transmission power of the node instead of the sum of distance used in WCA in order to elect the cluster head node which covers the largest range, thus minimizing the number of clusters generated. IWCA considers the average relative velocity instead of velocity and average distance instead of sum of distance with reference to WCA.WCA and its derived algorithms provide better performance with compromised set up delay. Introduction of more parameters result in increased set up delay.

Similarly, many such weighted algorithms are proposed for electing a cluster head. Apart from algorithms, routing protocols such as CEDAR, CBRP, etc. improve the scalability as well as performance of MANETs.

Cluster Based Routing Protocol [13], a on demand source routing protocol, clubs nodes into clusters and decreases control overhead during route discovery.K-Hop Cluster Based Routing Protocol [14] improves on CBRP in both clustering techniques and routing algorithm. It enhances theweighted clustering algorithm (WCA)for the election of Cluster Head.

InLocation-Aided Routing (LAR) [15] protocol the overhead of route discovery is decreased by utilizing location information of mobile nodes. Using GPS[16] for locationinformation, LAR protocol reduces the search space for a desired route. Reducing the search space results in fewer route discovery messages.By contacting a location service provider which knows the positions of all the nodes, the source node should first get the position of the destination mobile node when it wants to send data packets to a destination.

To localize the ad hoc network a wide variety of routing protocols [17-19] have been proposed over the years. Some techniques use GPS but for very few nodes. These nodes are often referred as anchor nodes or reference nodes. 'Completely GPS Free Localization [20-23] or 'Using Very Few Anchor Node' [24, 25] are the two types of localization approaches that provide techniques to localize the network in a GPS Less or GPS-Scarce area (LACBER). The GPS-less localization [26] approaches establish a virtual coordinate system and try to localize the network in that coordinate System. On the basis of distance measurement (using ToA or AoA or RSSI) or hop count these coordinate systems are established**.** **U**sing the above coordinate systems the exact location of the node cannot be determined due to absence of GPS.

Location Aided Cluster Based Energy-efficient Routing (LACBER) [27] is a location aided routing protocol proposed for GPS scarce ad hoc networks. In the network only a few nodes are GPS enabled and are capable of finding their own location using GPS.A few special nodes are equipped with antennas which can measure RSSI and the angle of arrival (AOA) of received signals from other nodes. The rest of the network can find their positions in a process using either GPS enabled or special nodes.

The LACBER protocol requires that each cluster must have at least one GPS enabled node or antenna equipped node in it. Compared to other cluster based routing protocols[28] the formation of clusters in LACBER protocol results in high control overhead. Using LACBERprotocol, determining the location of normal nodes with high mobility is a constraint.

3 Proposed Protocol

This paper proposes a ILCRP protocol where all the nodes in all the clusters are GPS enabled compared to a few nodes in a cluster as in LACBER protocol. The proposed protocol makes use of clusters as well as location information intensively. The exact information of the nodes is known to each other with the help of GPS. The protocol is divided into three phases. First phase is cluster formation followed by cluster maintenance. The last phase is route discovery phase.

In the proposed ILCRP protocol, the control overhead becomes less for route discovery due to its GPS capability. The proposed protocol delivers the packets more accurately with less end to end delays since the exact location of the source as well as destination nodes are known to respective cluster heads. Besides, the overhead decreases due to exact location information of the nodes at all cluster heads.

3.1 Cluster Formation

Clusters are formed between nodes which are m-hops away from the cluster head. All the nodes start in undecided stage. Since all the nodes are GPS enabled all the nodes can become cluster head. Initially all the nodes in the network broadcast a HELLO message with node ID and location information. Location information is obtained using GPS utility with location error tolerance limit of e. Let node ID be the MAC address as stated in FWCA. Based upon its updated neighbor node's list, each node computes its node value based on the following parameters:

- The degree differenceΔi: defined as the difference between the cluster's size 'N' and the actual number of neighbors. It allows estimating the remaining number of nodes that each node can still handle.

  Δi = |**di - N**| where d_i is the degree of the node and N is the threshold for number of nodes in the cluster
- The mobility of the node M.

Mobility of the node at time t2 is calculated using the below formula:

$$\boldsymbol{M} = \frac{\mathbf{1}}{(\boldsymbol{t2}-\boldsymbol{t1})}\left(\sqrt{(\boldsymbol{x2}-\boldsymbol{x1})^{\mathbf{2}}+(\boldsymbol{y2}-\boldsymbol{y1})^{\mathbf{2}}}\right)$$

Where $x1\ ,y1$ and $x2,y2$ are the co ordinates of the node at time $T1$ and $T2$respectively.

- The remaining battery power Pa of the node

Node Value=$\boldsymbol{W1} * \Delta \boldsymbol{i}\ +\ \ \boldsymbol{W2} * \boldsymbol{M}\ + \boldsymbol{W3} * \boldsymbol{Pa}$

Where $W1, W2, W3$ are the weights used and are in a relation such that$W1 + W2 + W3 = 1$.

After finding its value, all the nodes broadcast node value using a INFO message to its m-hop neighbors. Depending upon the node values, the node with the highest node value elects itself as Cluster Head by sending CH_INFO. In figure 1 the cluster C1 has one cluster head, one gateway node and four member nodes.

All other nodes store node IDs, location information, and its node values in its neighbor tables. Nodes usecluster adjacency table for inter cluster routing. Cluster head stores the adjacent cluster heads IDs, and uses gateway node IDs to reach them whereas member nodes store NULL value.

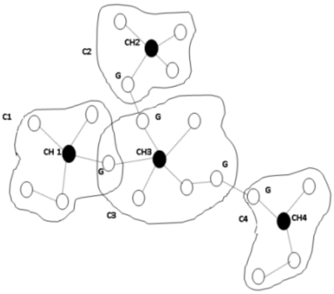


Fig. 1. ILCRP Cluster formation

3.2 Cluster Maintenance

The clusters have to be reorganized and reconfigured dynamically due to themobility of nodes in the ad hoc network. There are three major scenarios in a clusterfor reconfiguration. The sceneries are:

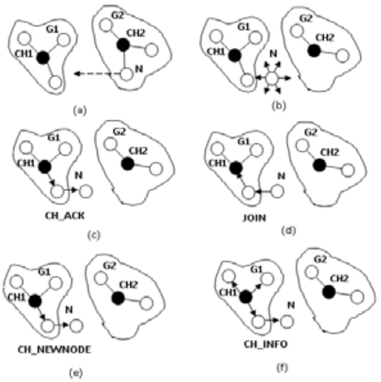
- Reduction in the Node Value of the Cluster Head
- Mobility of a Node
- Mobility of Cluster Head

3.2.1 Reduction in the Node Value of the Cluster Head

The cluster headdetermines its node value from time to time. When its node value falls belowthreshold value, the cluster head sends CH_RELEIVE to all its G-nodes in its cluster. After receiving CH_RELEIVE, all the nodes calculate the respective node values andconveys them to the cluster head. Now the cluster head broadcasts CH_RACK containing the Node ID of the succeeding Cluster Head.

3.2.2 Mobility of a Node

When a node moves from one cluster to another, the state of the node becomes undecided and it floods HELLO messages containing its nodeID and location information. On receiving the HELLO message, the cluster head verifies whether it has reached the threshold limit of the number of nodes that the cluster can accommodate at a given time. If the threshold has not been reached, itacknowledges the new node with CH_ACK. The new node sends back JOIN with its node value. Cluster head replies with CH_NEWNODE and broadcasts CH_INFOwith updated neighbor node. Beyond the threshold limit, the cluster head replieswith negative acknowledgement CH_NACK to the new node. The new node then tries with other cluster heads. It is explained in the figure2.



(a)Node N moves out of the cluster with CH2 (b) Node N broadcasts HELLO
(c) Node N receives CH_ACK from CH1 (d) Node N replies with JOIN
(e) CH replies Node N with CH_NEWNODE (f) CH1 broadcasts CH_INFO

Fig. 2. Mobility of a node

3.2.3 Mobility of Cluster Head

When the cluster head moves away from the farthest node in the cluster, the farthest node waits for HELLO messages after aperiod of refresh time T_{ref}.If the node receives the message, it still maintains the member state of the cluster. If it does not receive, it goes to undecided state. In theundecided state, it floods the neighboring nodes with HELLO messages indicating its presence. Upon receiving the acknowledgement from any reachable Cluster Head orany other nodes in a m-hop cluster, it replies with its INFO Message packet. Any reachable cluster head replies with its neighbor table and updates all the members in the cluster about the new node or else the previous cluster head updates the neighbortable after every T_{ref} and informs all the nodes.

3.3 Route Discovery

The route discovery is done using source routing in cluster based routing protocols, whereas in ILCRP protocol it is done using location information. So control overhead becomes extremely high in cluster based routing protocols compared to location based routing protocols for routing. Now, there are two instances of route discovery.

Table 1. Summarizes the messages used for formation as well as maintenance of the clusters

Message	Description
HELLO	Contains node ID, location information
INFO	Contains node value
CH_INFO	Cluster head Node ID , Cluster neighbor table
CH_ACK	Cluster Head acknowledges the new node's HELLO
JOIN	New Node requests to join the cluster after CH_ACK
CH_NEWNODE	Cluster Head acknowledges the new node's JOIN
CH_NACK	Cluster Head rejects the new node's HELLO
CH_RELIEVE	Notifies the members about its intention to resign as Cluster head
CH_RACK	Relieves finally after broadcasting new cluster head

3.3.1 Intra Cluster Routing

In intra cluster routing, all the node's GPS utility is made to sleep for reduced power consumption. All nodes in a cluster know about the location of other nodes in its cluster. So the source node forwards to the receiver node using the location information. If the destination node is one hop away from the receiver node, then source node sends the packet towards the destination node either using cluster head or using another node as shown in figure 3.

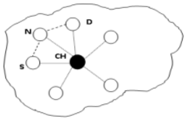


Fig. 3. Intra Cluster Routing Algorithm

When there is mobility of a node inside a cluster for a multi hop cluster, the use of LAR protocol results in higher efficiency. From fig 4, Node D moves with an average speed of v m/s from known location at $t0$.All the messages are routed to node D through node 1 at $t0$. After a time interval$t$, the node D is expected to be at a radialdistance of $Rd = v * t$ units from the location at t0. As shown in the figure 4, Node D is not reachable via node N1.Using LAR,expected region is reachable via node N2.

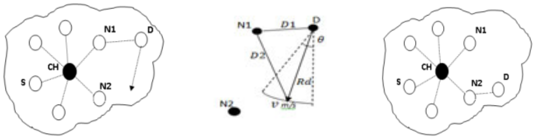
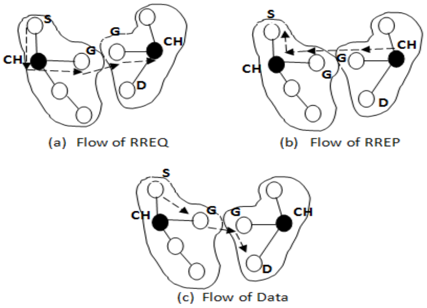


Fig. 4. Intra Cluster routing

3.3.2 Inter Cluster Routing

Using Cluster adjacency table, the cluster head sends an inter-cluster routing request packet(RREQ) to its gateway nodes to obtain routing information between clusters. Once the location information of the destination node is obtained using route reply packet(RREP), the source node sends the packet directly to gateway node as shown in the figure 5.



(a) Flow of RREQ (b) Flow of RREP

(c) Flow of Data

Fig. 5. Inter Cluster Routing

4 Simulation Results

4.1 Simulation Parameters

- Performed using NS-2 network simulator[29] with MANET extensions.
- IEEE 802.11 is used as the MAC layer protocol.
- The radio model simulates with a nominal bit rate of 2Mbps.
- Nominal transmission range of 125 meters.
- The radio propagation model is the two-ray ground model.
- First 100 nodes are deployed for one experiment and then 100 nodes are used for another experiment in a field of 1000m X 1000m.
- The traffic pattern is CBR (constant bit rate) with a network traffic load of 4 packet/seconds and the packet lengths are all 512 bytes.
- The mobility model used is the Random Waypoint Model
- The pause time of the node mobility as the independent variable reflects the degree of the node mobility. The small pause time means intense node mobility and large pause time means slow node mobility. The pause time is maintained as 5 seconds.
- The simulation time is 600 seconds.
- The simulations are performed by varying the speed from 2 m/s to 10 m/s at its successive stages with an interval of 2 m/s for a maximum of 40 nodes.
- The simulations are performed by creating 20, 40, 60, 80, 100 nodes at its successive stages.

4.2 Performance Metrics

For evaluating the performance of ILCRP, the metrics chosen are Packet Delivery ratio, Control Overhead, and End to End Delay.

4.2.1 End to End Delay

End to End Delay indicates the time lapse between the source and destination nodes in the network. Figure 6 and Figure 7 shows that the end to end delay reduces if the exact locations of all the nodes are obtained. On increasing the mobility of the nodes, the delay increases due to reconfiguration of the clusters. The end to end delay also increases due to increase in the number of nodes due to more number of hops.

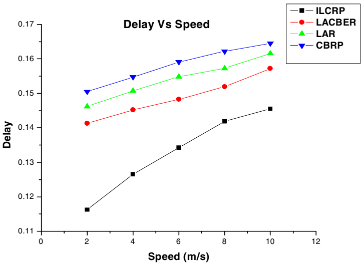


Fig. 6. Delay Vs Speed

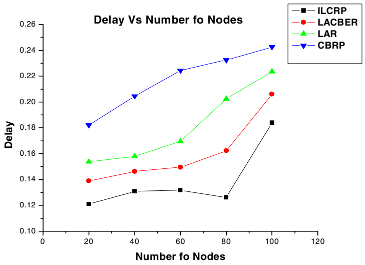
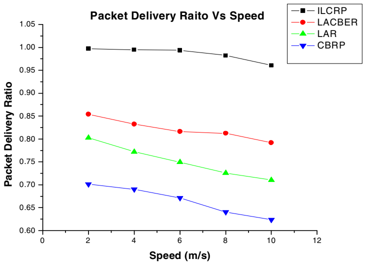


Fig. 7. Delay Vs Number of Nodes

4.2.2 Packet Delivery Ratio

It is defined as the ratio of total number of packets that have reached the destination node to the total number of packets originated at the source node. The location information of the nodes make the packets route, loop free which results in high packet delivery ratio. On increasing the mobility or speed of the nodes, the delivery ratio decreases since most of the nodes move away from each other. Increasing the number of nodes decreases the delivery ratio due to tightly coupled clusterconfiguration. Figure 8 and Figure 9 confirms the packet delivery ratio between ILCRP and LACBER, LAR, CBRP.



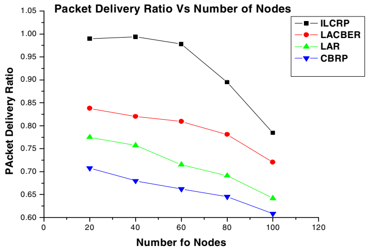


Fig. 8. Packet Delivery Ratio Vs Speed **Fig. 9.** Packet Delivery Ratio Vs Number of Nodes

4.2.3 Control Overhead

It is defined as the ratio of the number of control packets transmitted to the number of the data packets delivered. Usage of cluster based routing protocol for clustering and exact location information for route discovery reduces the control overhead in the network. Figure 10 and Figure 11 shows the control overhead ratio between ILCRP, LACBER,LAR and CBRP. It increases when the mobility of the nodes as well as number of nodes increases.

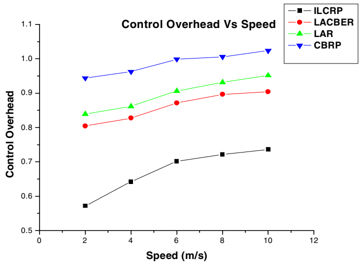


Fig. 10. Control Overhead Vs Speed

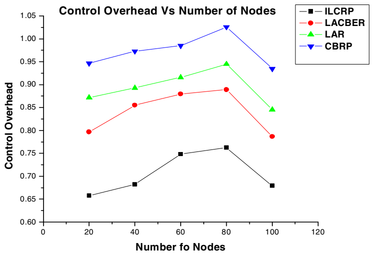


Fig. 11. Control Overhead Vs Number of Nodes

5 Conclusion

A detailed performance evaluation of ILCRP has been presented in the paper. Number of nodes as well as mobility of nodes is varied for the simulation using NS2.The exact location information of nodes in ILCRP increases the delivery ratio and makes the route, loop free. Location information of all the nodes keeps the exchange information as well as end to end delay very low in ILCRP compared to other protocols. On increasing the mobility of the nodes as well as the number of nodes, the overall performance of ILCRP has been found declining due to reconfiguration of the network. Our simulation results using NS2 have shown that GPS enabled MANETs perform better than GPS free as well as GPS Scarce MANETs.

References

1. Iwata, A., Chiang, C., Pei, G., Gerla, M., Chen, T.: Scalable Routing strategies for AdHoc Wireless Networks. IEEE J. Select. Areas Communication 17(8), 1369–1379 (1999)
2. Ephremides, A., Wieselthier, J.E., Baker, D.J.: A design concept for reliable mobile radio networks with frequency hoping signaling. Proc. IEEE 75(1), 56–73 (1987)
3. Gerla, M., Tsai, J.T.C.: Multicluster, mobile, multimedia radio network. Wirel. Netw. 1, 255–265 (1995)
4. Parekh, A.K.: Selecting routers in ad hoc wireless network. In: Proceedings of the SBT/IEEE International Telecommunication Symposium, ITS (August 1994)
5. Basu, P., Khan, N., Little, T.D.C.: A mobility based metric for clustering in mobile ad hoc networks. In: Proceedings of IEEE ICDCS 2001 Workshop on Wireless Networks and Mobile Computing, Phoenix, AZ (2001)
6. Basagni, S.: Distributed clustering for ad hoc networks. In: Proceedings of International Symposium on Parallel Architectures, Algorithms and Networks, pp. 310–315 (June 1999)
7. Das, S., Chatterjee, M., Turgut, D.: WCA: a weighted clustering algorithm for mobile ad hoc networks. J. Cluster Comput. 5(2), 193–204 (2002)
8. Das, S., Chatterjee, M., Turgut, D.: An on demand weighted clustering algorithm(WCA) for ad hoc networks. In: Proceedings of IEEE GLOBECOM 2000, San Francisco, pp. 1697–1701 (November 2000)
9. Chatterjee, M., Das, S.K., Turgut, D.: A weight based distributed clustering algorithm for mobile ad hoc networks. In: Prasanna, V.K., Vajapeyam, S., Valero, M. (eds.) HiPC 2000. LNCS, vol. 1970, pp. 511–521. Springer, Heidelberg (2000)

10. Ghosh, R., Basagni, S.: Limiting the impact of mobility on ad hoc clustering. In: Proceedings of the 2nd ACM International Workshop PE-WASUN 2005, Montreal, CA, pp. 197–204 (2005)
11. Zhao, X., Gu, X., Sun, Z., Ren, C.: An Intelligent Weighted Clustering Algorithm(IWCA) for Ad Hoc. In: Software Engineering, WCSE 2009 (2009)
12. Tahiti, French Polynesia.: A Flexible Weight Based Clustering Algorithm in Mobile Ad hoc Networks.In: International Conference on Systems and Networks Communication ICSNC 2006 (2006)
13. Sucec, J., Marsic, I.: Clustering overhead for Hierarchical Routing in Mobile Ad Hoc Networks. In: IEEE Infocom (2002)
14. Nanjing: A Multi-Hop Cluster Based Routing Protocol for MANET.In: First International Conference on Information Science and Engineering (2009)
15. Ko, Y.-B., Nitin, H., Vaidya: Location-Aided Routing (LAR) in Mobile Ad Hoc Networks. In: Proceedings of the 4th annual ACM/IEEE international conference on Mobile computing and networking, October 25-30, pp. 66–75 (1998)
16. Parkinson, B., et al.: Global Positioning System: Theory and Application. Progress in Astronautics and Aeronautics 1, 163 (1996)
17. Mauve, M., Widmer, J., Hartenstein, H.: A Survey on Position-Based Routing in Mobile Ad Hoc Networks. IEEE Network Magazine 15, 30–39 (2001)
18. Tomar, G.S., Member IEEE Tomar, R.S.: Position Based Routing for Mobile Ad Hoc Networks. In: Second UKSIM European Symposium onComputer Modeling and Simulation
19. Mohammad, A.M.: Energy efficient Location Aided Routing Protocol for Wireless MANETs. International Journal of Computer Science and Information Security 4(1&2) (2009)
20. Capkun, S., Hamdi, M., Hubaux, J.P.: GPS free positioning in Mobile Ad Hoc Networks. Cluster Computing Journal 5(2), 157–167 (2002)
21. Caruso, A., Chessa, S., De, S., Urpi, R.: GPS free coordinate assignment and routing in Wireless Sensor Networks. In: Proceedings of the IEEE INFOCOM, pp. 150–160 (2005)
22. Akcan, H., Kriakov, V., Bronnimann, N.: GPS-Free node localization in Mobile Sensor Networks. In: Proceedings of The 5thACM International Workshop on Data Engineering for Wireless and Mobile Access, pp. 35–42 (2006)
23. Iyenger, R., Sikdar, B.: Scalable and Distributed GPS free positioning for Sensor Networks. In: Proceedings of IEEE conference on communication ICC 2003, vol. 1, pp. 338–342 (2003)
24. Kwon, O.-H., Song, H.-J.: Counting-Based Distance Estimations and Localizations in Wireless Sensor Networks. In: Gavrilova, M.L., Gervasi, O., Kumar, V., Tan, C.J.K., Taniar, D., Laganá, A., Mun, Y., Choo, H. (eds.) ICCSA 2006. LNCS, vol. 3983, pp. 306–315. Springer, Heidelberg (2006)
25. Chu, H.-C., Jan, R.-H.: A GPS-less, outdoor, self positioning method for Wireless Sensor Networks. In: Ad Hoc Networks, vol. 5(5), pp. 547–557. Elsevier Science, Amsterdam (2007)
26. Bulusu, N., Heidemann, J., Estrin, D.: GPS-less Low Cost Outdoor Localization For Very Small Devices. IEEE Personal Communications Magazine, Special Issue on Smart Spaces and Environments (October 2000)
27. Deb, D., Roy, S.B., Chaki, N.: LACBER: A New Location Aided Routing Protocol for GPS Scarce MANET. International Journal of Wireless & Mobile Networks (IJWMN) 1(1) (August 2009)
28. Jahanbakhsh, S.K., Hajhosseini, M.: Improving Performance of Cluster Based Routing Protocol using Cross-Layer Design
29. The Network Simulator ns-2, Information Sciences Institute, USA.Viterbi School of Engineering (September 2004), `http://www.isi.edu/nsnam/ns/`

State of the Art in Wireless Sensor Networks Operating Systems: A Survey

Muhammad Omer Farooq[1], Sadia Aziz[2], and Abdul Basit Dogar[3]

[1] Transmission Systems Research Group, Jacobs University Bremen, Germany
[2] Computer Engineering Department, CASE, Islamabad, Pakistan
[3] Department of Computer Science, Virtual University of Pakistan, Lahore, Pakistan
m.farooq@jacobs-university.de, sadia.aziz@case.edu.pk,
abasit126@yahoo.com

Abstract. This paper, presents a survey on current state of the art in Wireless Sensor Networks (WSNs) Operating Systems (OSs). WSN is composed of miniature senor and resource constraint devices. WSN is highly dynamic network because nodes die out due to severe environmental conditions and battery power depletion. Stated characteristics of WSN impose additional challenges on OS design for WSN. Therefore; OS design for WSN deviates from traditional OS design. The purpose of this survey is to point out strengths and weaknesses of contemporary OS for WSNs, keeping in view the requirements of emerging WSNs applications. State of the art, in operating systems for WSNs has been examined in terms of Architecture, Scheduling, Threading Model, Synchronization, Memory Management, and Communication Protocol support. The examination of these features is performed for both real time and non real time operating systems for WSNs. We believe that this survey will help the network designers and programmers to choose the right OS for their network and applications. Moreover, pros and cons of different operating systems will help the researchers to design more robust OSs for WSNs.

Keywords: Wireless Sensor Networks (WSN), Operating System (OS), Embedded Operating Systems.

1 Introduction

Advances in Micro-Electro Mechanical System (MEMS) based sensor technology has led to the development of miniaturized and cheap sensor nodes, capable of wireless communication, sensing and performing computations. Wireless Sensor node is composed of micro-controller, transceiver, timer, memory and analog to digital converter. Figure 1 shows the block diagram of a typical sensor node. Sensor nodes are deployed to monitor multitude of natural and unnatural phenomenon i.e., habitant monitoring, wild life monitoring, patient monitoring, industrial process monitoring, battle field monitoring, traffic control, home automation to name a few. Sensor nodes have constraint resources i.e., small amount of battery, few kilobytes of memory and a microcontroller that operates at very low frequency compared to traditional contemporary processing units. These resource constraint tiny sensors are an example

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 616–631, 2010.

of System on Chip (SoC). Dense deployment of sensor nodes in sensing field and distributed processing through multi-hop communication among sensor nodes is required to achieve high quality and fault tolerance in WSN. Application areas for sensors are growing and new applications for sensor networks are emerging rapidly.

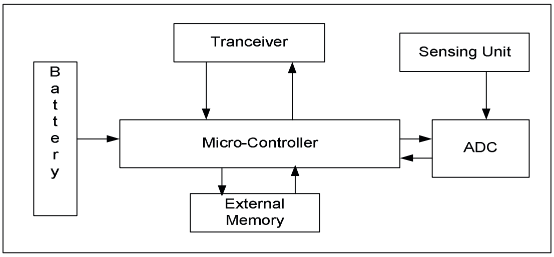


Fig. 1. Sensor Node Architecture

OS acts as a resource manager for complex systems. In a typical system these resources include processors, memories, timers, disks, mice, keyboard, network interfaces etc. The job of OS is to manage allocation of these resources to users in an orderly and controlled manner. An OS multiplex system resources in two ways i.e., in time and in space. Time multiplexing involves different programs taking turn in using the resources. Space multiplexing involves different programs getting part of the resource possibly at the same time.

Literature exist that survey the application, transport, network and Medium Access Control (MAC) protocols for WSN, one such effort was made in [11]. A survey on Operating Systems for WSN also exists and published in [10]. Since [10] was published, new features like Architecture, Execution Model, Reprogramming, Scheduling and Power Management have been introduced in contemporary operating systems for WSN, hence the need for this survey remains important.

In this survey, we have examined the core OS features like Architecture, Scheduling, Threading Model, Synchronization, Memory Management, and Communication Protocol Support in both real time and non-real time WSN operating systems. We have discussed different design approaches taken by different WSN operating systems with their relative pros and cons.

Section 2 presents major design concern for WSNs OS. TinyOS has been investigated in Section 3. In Section 4, we investigate the Contiki operating system and Section 5 presents MANTIS operating System. We have identified future research directions in Section 6. Finally, this paper is concluded in Section 7.

2 Major Concerns in WSN OS Design

This section, gives details of major issues related to an OS design for WSN.

2.1 Architecture

Organization of an OS constitutes its structure. In an OS perspective, architecture of the OS kernel makes up its structure. Architecture of an OS has an influence on the size of the OS kernel as well as on the way it provides services to the application programs. Some of the well known OS architectures are Monolithic Architecture, Micro-Kernel Architecture, Virtual Machine Architecture and Layered Architecture.

A monolithic architecture in-fact does not have any structure. Services provided by an OS are implemented separately and each service provides an interface for other services. Such architecture, allows bundling of all the required service together into a single system image thus, results in larger OS footprint. Advantage of monolithic architecture is that cost of module interaction is low. Disadvantages associated with this architecture are: system is hard to understand, modify, and to maintain. Disadvantages associated with monolithic kernels make them as a bad OS design choice for sensor nodes.

Alternate choice for an OS design is Microkernel architecture. In microkernel minimum functionality is provided inside the kernel. Thus, kernel size is significantly reduced. Most of the OS functionality is provided in user level servers like file server, memory server, time server etc. If one server crashes down whole system does not crash. Microkernel kernel architecture provides better reliability, ease of extension and customization. The disadvantage associated with microkernel is poor performance because of user kernel boundary crossing. Microkernel is the design choice for many embedded OS due to the small kernel size secondly; context switches are far less in embedded systems. Thus, boundary crossing is less compared to traditional systems.

Virtual machine architecture is another design alternative for an OS. Main idea is to export virtual machines to user programs, which resemble hardware. A virtual machine has all hardware features. Advantage is portability. Disadvantage is poor system performance.

Layered OS architecture implements services in the form of layers. Advantages associated with layered architecture are: more manageability, understandability, and reliability. The disadvantage is that it is not flexible.

An OS for Wireless Sensor Network should have an architecture that results in a small kernel footprint. Architecture must allow extensions to the kernel if required. Architecture must be flexible i.e., only application required services get loaded onto the system.

2.2 Resource Sharing

Responsibility of an OS includes resources allocation and resource sharing is of immense importance when multiple programs are concurrently executing. Majority of sensor network OS now provide some sort of multithreading therefore, there must be a resource sharing mechanism available. This can be performed in time e.g., scheduling of a process on the CPU and in space e.g., writing data to system memory. In some cases we need a serialized access to resources and this is done through the use of synchronization primitives.

2.3 Protection

In traditional operating system, protection refers to protecting of one process from another. In early sensor network operating systems like TinyOS [1] there was no memory management available. Early, operating systems for sensor networks assumed that only a single thread executes on a process therefore; there is no need for memory protection. Latest WSN involves multiple threads of execution therefore; memory management becomes an issue for WSN OS.

2.4 Performance

In OS, how we make the system all go fast is called performance. Performance of a system can be measured by throughput, access time, and response time. The ultimate responsibility of an OS designer is to enhance the overall performance of the system, keeping in view the kind of application that runs on the system.

2.5 Communication

In OS context, communication refers to inter-process communication within the system as well as with other nodes in the network. Sensor networks operate in a distributive environment therefore; senor nodes communicate with other nodes in the network. The job of sensor network OS is to provide Application Programming Interface (API) that provides easy and energy efficient way of communication. It is possible that sensor network is composed of heterogeneous sensor nodes therefore; communication protocol provided by the OS must also consider heterogeneity. In network based communication, OS should provide transport layer, network layer and MAC layer protocol implementation.

2.6 Scheduling

Central Processing Unit (CPU) scheduling determines the order in which tasks are executed on CPU. In traditional computer systems, the goal of a perfect scheduler is to minimize latency, maximize throughput, maximize resource utilization, and fairness.

The type of scheduling algorithm for sensor network typically depends on the nature of the application. For applications having real time requirements we need to use real time scheduling algorithm and for other applications we can use non-real time scheduling algorithms.

Sensor networks are being used in both real time and non-real time phenomenon therefore, a sensor network OS must provide scheduling algorithm that can accommodate the application requirements. Moreover, scheduling algorithm should be memory and energy efficient.

2.7 Multithreading

Multithreading provides a convenient application development environment. In threaded systems, context switching and scheduling are the source of major overhead [8]. We know that sensor nodes are battery operated, memory limited and have low computational power. Therefore, sensor network operating system should support high concurrency with minimal memory usage and low energy consumption.

3 TinyOS

TinyOS [1] is an open source flexible component based, and application specific operating system designed for sensor networks. TinyOS can support concurrent programs with very low memory requirements. The OS has footprint that fits in 400 bytes. TinyOS component library includes network protocols, distributed services, sensor drivers, and data acquisition tools. Following subsections survey TinyOS design in more detail.

3.1 Architecture

TinyOS fall under the monolithic architecture. TinyOS uses the component model and according to the requirements of an application different components are glued together with the scheduler to compose a static image that runs on the mote. A component is an independent computational entity that exposes one or more interfaces. Components have three computational abstractions: commands, events, and tasks. Mechanisms for inter-component communication are commands and events. Tasks are used to express intra-component concurrency. A command is a request to perform some service, while the event signals the completion of the service.

TinyOS provides single shared stack and there is no separation between kernel space and the user space. For program execution TinyOS uses an event driven model. Following figure shows the TinyOS architecture.

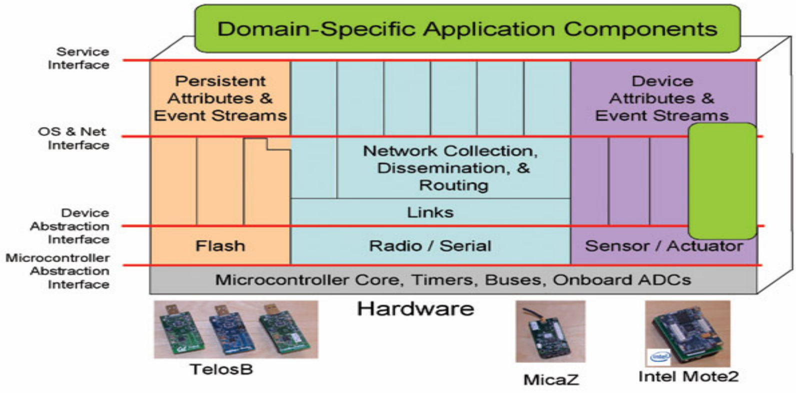

Domain-Specific Application Components
Service Interface
Persistent Attributes & Event Streams
Device Attributes & Event Streams
OS & Net Interface
Network Collection, Dissemination, & Routing
Links
Device Abstraction Interface
Flash
Radio / Serial
Sensor / Actuator
Microcontroller Abstraction Interface
Microcontroller Core, Timers, Buses, Onboard ADCs
Hardware
TelosB
MicaZ
Intel Mote2

Fig. 2. TinyOS Architecture

3.2 Scheduling

Earlier versions of TinyOS support non preemptive First In First Out (FIFO) scheduling algorithm. Therefore; those versions of TinyOS does not support real time application. This prevents TinyOS usage in real time systems. The core of the execution model in TinyOS is task that runs to completion in FIFO manner. Since, TinyOS supports only non preemptive scheduling therefore, task must obey run to completion semantics. Tasks run to completion with respect to other task but they are not atomic with respect to interrupt handlers and commands and events they invoke. Since TinyOS uses FIFO scheduling therefore, disadvantages associated with FIFO scheduling are also associated with the TinyOS scheduler. The wait time for a task depends on the tasks arrival time. FIFO scheduling can be unfair to latter tasks especially when short tasks are waiting behind the longer ones.

In [1], authors have claimed that they have added support for Earliest Deadline First (EDF) scheduling algorithm in TinyOS, to facilitate real time application. EDF scheduling algorithm does not produce a feasible schedule when tasks content for the resources. Thus, TinyOS does not provide a good real time scheduling algorithm if different threads content for resources.

3.3 Threading Model and Synchronization

Earlier versions of TinyOS do not provide any multithreading support. TinyOS version 2.1 provides support for multithreading and these TinyOS threads are called TOS Threads. In [3], authors pointed out the problem that given the motes resource constraints, an event based OS permits greater concurrency. However, preemptive threads offer an intuitive programming paradigm. TOS threading package provides ease of a threading programming model with the efficiency of an event driven kernel. TOS threads are backward compatible with existing TinyOS code. TOS threads use cooperative threading approach, i.e., TOS threads rely on applications to explicitly yield the processor. This adds on an additional burden on the programmer to manage the concurrency explicitly. Application level threads in TinyOS can preempt other application level threads but they cannot preempt tasks and interrupt handlers. High priority kernel thread is dedicated to run the TinyOS scheduler. For communication between the application threads and kernel, TinyOS 2.1 provides the mechanism of message passing. When an application program makes a system call, it does not directly execute the code rather it posts a message to the kernel thread by posting a task. Afterwards, kernel thread preempts the running thread and executes the system call. This mechanism ensures that only kernel directly executes TinyOS code. System calls like *Create, Destroy, Pause, Resume and Join* are provided by the TOS threading library.

TOS threads dynamically allocate Thread Control Blocks (TCB) with space for fixed size stack that does not grow over time. TOS Threads context switches and system calls introduce an overhead of less than 0.92% [3].

Earlier versions of TinyOS impose atomicity by disabling the interrupts i.e., telling the hardware to delay handing the external events until aftersystem is done with the atomic operation. This scheme works well on uni-processor systems. Secondly, critical section can occur in the user level threads and the designer of OS does not want user to disable the interrupts due to system performance and usability issues. This problem is circumvented in TinyOS version 2.1. It provides synchronization support with the help of condition variables and mutexes. These synchronization primitives are implemented with the help of special hardware instructions e.g., test & set instruction.

3.4 Memory Management and Safety

In [2], efficient memory safety for TinyOS is presented. In sensor nodes, hardware based memory protection is not available and the resources are scarce. Resource constraints necessitate the use of unsafe, low level languages like nesC [17]. In TinyOS version 2.1 memory safety is incorporated. The goals for memory safety as given in [2] are: trap all pointer and array errors, provide useful diagnostics, and provide recovery strategies. Implementations of memory safe TinyOS exploits the concept of Deputy. Deputy is a resource to resource compiler that ensures type and memory safety for C code. Code compiled by Deputy relies on a mix of compile and run time checks to ensure memory safety. Safe TinyOS is backward compatible with earlier version of TinyOS. Safe TinyOS tool chain inserts checks into the application code to ensure safety at run time. When a check detects that safety is about to be violated, code inserted by Safe TinyOS take remedial action.

3.5 Communication Protocols Support

Earlier versions of TinyOS use two multi-hop protocols: dissemination [14] and TYMO [15]. Dissemination protocol, reliably delivers data to every node in the network. This protocol enables administrators to reconfigure query and reprogram a network. Dissemination Protocol provides two interfaces: DisseminationValue and DisseminationUpdate. A producer should call the DisseminationUpdate. The command DisseminationUpdate.chage() should be called each time the producers wants to disseminate a new value. On the other hand DisseminationValue is for the consumer. The event DisseminationValue.changed() is signaled each time the dissemination value s changed. TYMO is the implementation of the DYMO protocol, a routing protocol for mobile ad hoc networks. In TYMO, packet formats have changed and it has been implemented on top of the active messaging stack.

K. Lin et al [16], presents DIP a new dissemination protocol for sensor networks. DIP is a data discovery and dissemination protocol that scales to hundreds of values. At MAC layer TinyOS provide implementation of the following protocols: a single hop TDMA protocol, a TDMA/CSMA hybrid protocol which implements Z-MAC's slot stealing optimization, and an optional implementation of 802.15.4 complaint MAC is available.

4 Contiki

Contiki [5], is a lightweight open source OS written in C language for WSN sensor nodes. Contiki is highly portable OS and it is build around an event driven kernel. Contiki provides preemptive multitasking that can be used at the individual process level. A typical Contiki configuration consumes 2 kilobytes of RAM and 40 kilobytes of ROM. A full Contiki installation includes features like: multitasking kernel, preemptive multithreading, proto-threads, TCP/IP networking, IPv6, Graphical User Interface, web browser, personal web server, simple telnet client, screensaver, and virtual network computing. In the following subsections, we shall explore Contiki OS in more detail.

4.1 Architecture

Contiki OS, follows the hybrid architecture i.e., it combines advantages of events and threads. At the kernel level it follows the event driven model but it provides optional threading facility to individual processes. Contiki kernel comprises of a lightweight event scheduler that dispatches events to the running processes. Process execution is triggered by the events dispatched by the kernel to the processes or by the polling mechanism. Polling mechanism is used to avoid race conditions. Any scheduled event will run to completion however, event handlers can use internal mechanism for preemption.

There are two kinds of events supported by Contiki OS: asynchronous events and synchronous events. The difference between two is that synchronous events are dispatched immediately to the target process that causes it to be scheduled, on the other hand asynchronous events are more like deferred procedure calls that are enqueued and dispatched later to the target process.

Polling mechanism used in Contiki can be seen as high priority events that are scheduled in between each asynchronous event. When a poll is scheduled all processes that implement a poll handler are called in order of their priority.

All OS facilities e.g., senor data handling, communication, and device drivers are provided in the form of services. Each service has its interface and implementation. Application using a particular service needs to know its interface. Application is not concerned about the implementation of a service. Following is the block diagram of Contiki OS architecture, as given in [18].

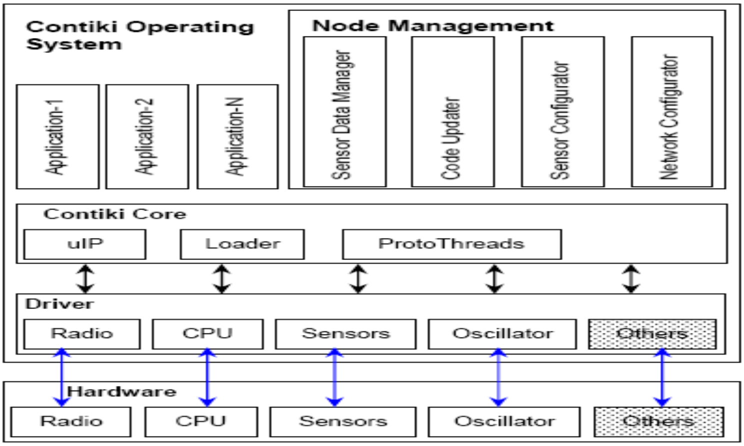


Fig. 3. Contiki Architecture [18]

4.2 Scheduling

Contiki is an event driven OS therefore, it does not employ any sophisticated scheduling algorithm. Events are fired to the target application as they arrive. In case of interrupts, interrupt handlers of an application runs w.r.t. its priority.

4.3 Threading Model and Synchronization

Contiki supports preemptive multithreading. Multi-threading is implemented as a library on top of the event driven kernel. The library can be linked with the applications that require multithreading. Contiki multithreading library is divided in two parts: a platform independent part and a platform specific part. The platform independent part interfaces to the event kernel and the platform specific part of the library implements stack switching and preemption primitives. Since, preemption is supported therefore; preemption is implemented using the timer interrupt and the thread state is stored on a stack. Available threading system calls are: *mt_yield(), mt_post(id,event,dataptr), mt_wait(event,dataptr), mt_exit(), mt_start(thread,funptr,dataptr), mt_exec(thread)* .

For multithreading Contiki uses protothreads [19]. Protothreads are designed for severely memory constraint devices because they are stack less and lightweight. Main features of protothreads are: very small memory overhead only two bytes per protothreads, no extra stack for a thread, highly portable, can be used with or without OS, provides blocking wait without full multithreading and stack switching, and freely available under BSD like open source license.

Since, events run to completion and Contiki does not allow interrupt handlers to post new events therefore; there is no process synchronization provided in Contiki.

4.4 Memory Management

Contiki supports dynamic memory management apart from this it also supports dynamic linking of the programs. In-order to guard against memory fragmentation Contiki uses Managed Memory Allocator [22]. Contiki's managed memory allocator makes sure that memory fragmentation does not occur. The primary responsibility of managed memory allocator is to keep the allocated memory free from fragmentation by compacting the memory when blocks are free. Therefore, a program using the memory allocator module cannot be sure that allocated memory stays in place.

For dynamic memory management Contiki also provide memory block management functions [22]. This library provides simple but powerful memory management functions for blocks of fixed length. A memory block is statically declared using the MEMB() macro. Memory blocks are allocated from the declared memory by the memb_alloc() function, and are deallocated using memb_free() function.

4.5 Communication Protocol Support

Contiki supports rich set of communication protocols. In Contiki, we can use both versions of IP i.e., IPv4 and IPv6. Contiki provides the implementation of *u*IP TCP/IP protocol stack which makes it possible to communicate with TCP/IP protocol suite even on small 8 bit micro-controllers. *u*IP does not require its peers to have full size stacks, but it can communicate with peers running a similar lightweight stack.

*u*IP implementation have the minimum set of features needed for a full TCP/IP stack. *u*IP is written in C language, it can only support one network interface, and it supports TCP, UDP, ICMP, and IP protocols.

Since, memory is a scare resource in embedded devises therefore; *u*IP uses memory efficiently by using memory management mechanisms. *u*IP stack does not use explicit dynamic memory allocation. It uses a global buffer to hold the incoming data packets. Whenever, a packet is received Contiki places it in the global buffer and notifies the TCP/IP. If it's a data packet, TCP/IP notifies the appropriate application. Application needs to copy the data in the secondary buffer or it can immediately process the data. Once the application is done with the received data, Contiki overwrites the global buffer with new incoming data. If application delays data processing, then data can be overwritten by new incoming data packets.

Contiki provides implementation of RPL (IPv6 routing protocol for low power and lossy networks) [21] by the name ContikiRPL [20]. ContikiRPL operates on low power wireless links and lossy power line links.

5 MANTIS

MANTIS, MultimodAl system for NeTworks of In-situ wireless sensors provides a new multithreaded operating system for wireless sensor networks. MANTIS is a lightweight and energy efficient operating system and it has a footprint of 500 bytes, which includes kernel, scheduler, and network stack. MANTIS Operating System (MOS), key feature is that it is portable across multiple platforms i.e., we can test MOS applications on PDA, and on x86 personal computers afterwards, application can be ported to the sensor node. MOS also supports remote management of sensor nodes through dynamic programming. MOS is written in C and it supports application development in C. Following subsection discusses the design features of MOS in more detail.

5.1 Architecture

MOS follows the layered multithreading design as shown in Figure 4. In layered architecture, services provided by an OS are implemented in layers. Each layer acts as an enhanced virtual machine to the layers above. Following are the different services implemented at each layer of MOS.

Layer 3: Network Stack, Command Server, and User Level Threads
Layer 2: MANTIS system API
Layer 1: Kernel/Scheduler, Communication Layer (MAC and PHY), and Device Drivers
Layer 0: Hardware

Layering structure imposes a hierarchical structure and fixed layering is not flexible. Crossing a layering boundary has associated overheads. Due to the layered approach an OS gets more reliable, manageable, understandable, and easily modifiable. Since, senor networks OS are not that complex as compared with traditional OS therefore; it's not a bad idea to use layering architecture.

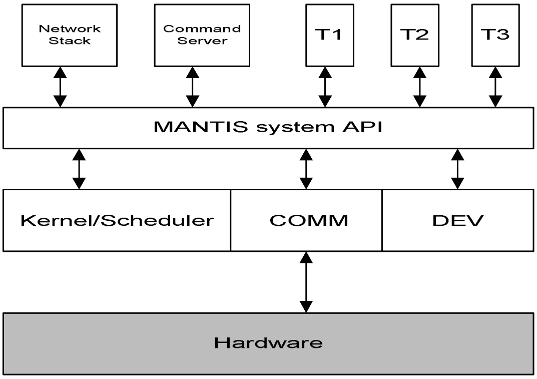


Fig. 4. MANTIS OS Architecture. Kernel Scheduler, COMM, DEV, MANTIS System API, Network Stack, and Command Server comprises MOS

MOS supports rich set of Application Programming Interface (API), written in C language. The choice of C language API simplifies cross platform support [4]. The C code developed for MANTIS sensor can be compiled to X86 PCs with little or no modification.

MOS kernel only handles the timer interrupt all other interrupts are directly sent to associated device drivers. When a device driver receives an interrupt, it posts a semaphore in order to activate a waiting thread, and this thread handles the event that has caused the interrupt.

5.2 Scheduling

MOS uses preemptive priority based scheduling. MOS uses a UNIX like scheduler with multiple priority classes and it uses round robin within each priority class. The length of time slice is configurable, by default it is set to 10 milliseconds (ms). The scheduler uses a timer interrupt for context switches. Context switches are also triggered by system calls and semaphore operations. Energy efficiency is achieved by the MOS scheduler by switching microcontroller to sleep mode when application threads are idle.

The ready queue of the MOS scheduler comprises of five priorities ranging from high to low: Kernel, Sleep, High, Normal, and Idle. The scheduler schedules the highest priority task in the ready queue. The task either runs to completion or gets preempted if its time slice expires. For time slicing MOs scheduler uses 16 bit timer. When there is no thread in the ready queue, system gets to the sleep mode. If the system is suspended on I/O, then the system enters the moderate idle sleep mode. If the application threads have called sleep system call, then system gets to deep power save sleep mode. A separate queue maintains the ordered list of thread that have called the sleep(), and is ordered by sleep time from low to high. The sleep priority in the ready queue enables newly woken threads to have the highest priority so that they can be serviced first after wake up.

The MOS kernel maintains ready list head and tail pointers for each priority level. There are 5 priority levels and these pointers consume 20 bytes in total. These two pointers helps in fast addition and deletion of threads from a ready queue hence, improved performance in manipulating thread lists. It also uses a current thread pointer of 2 bytes, an interrupt status byte, and one byte of flags. The total static overhead for scheduling is 144 bytes.

MOS scheduler uses round robin scheduling within the each priority class. This means threads of highest priority class can make lower priority class threads to strave. MOS use priority scheduling that may support real time task better than TinyOS scheduler. But it still needs real time schedulers like Rate Monotonic and Earliest Deadline First in-order to accommodate real time tasks.

5.3 Threading Model and Synchronization

MOS supports preemptive multitasking. MOS team designed a multithreaded OS because of the facts presented in [23], i.e., “A thread driven system can achieve the high performance of event based systems for concurrency intensive applications, with appropriate modification to the threading package.” Memory of the sensor node is a scare resource therefore, MOS maintains two logically distinct sections of RAM: the space for global variables that is allocated at the compile time, and the rest of the RAM is managed as a heap. Whenever a thread is created, stack space is allocated by the kernel out of heap. The stack space is returned to heap once the thread exit. Thread table is the main data structure that is being handled by the MOS kernel. In thread table, there is one entry per thread. MOS statically allocates memory for the thread table therefore, there can be fixed maximum number of threads hence fixed overhead. The maximum number of threads can be adjusted at the compile time. By default it is 12. Thread table entry comprises of 10 bytes and it contains: current stack pointer, stack boundary information (base pointer and size), pointer to thread starting function, thread priority level, and pointer to next thread. Once a thread is suspended its context is saved on the stack. Since, each thread table entry comprises of 10 bytes and by default 12 threads can be created therefore, associated overhead in terms of memory is 120 bytes. By default each thread gets a time slice of 10 ms and context switch happens with the help of timer interrupt. System calls and posting of a semaphore operation can also trigger context switch.

Multithreading support in MOS comes at the cost of context switching and stack memory overhead. In [4], the argument presented in favor of context switching overhead is that it is only a moderate issue in WSNs. It has been observed that each context switch incurs 60 microseconds overhead in comparison to this default time slice is much larger i.e., 10 ms which is less than 1% of the microcontroller cycles. Second cost is of stack memory allocation. The default thread stack in MOS is 128 bytes and MICA2 motes have a 4 KB RAM. Since, MOS kernel occupies 500 bytes therefore considerable space is available to support threading.

MOS avoids race conditions using binary mutex and counting semaphores. Semaphore in MOS is a 5 byte structure and it is declared by an application as needed. Semaphore structure contains a lock or count byte along with head and tail pointers.

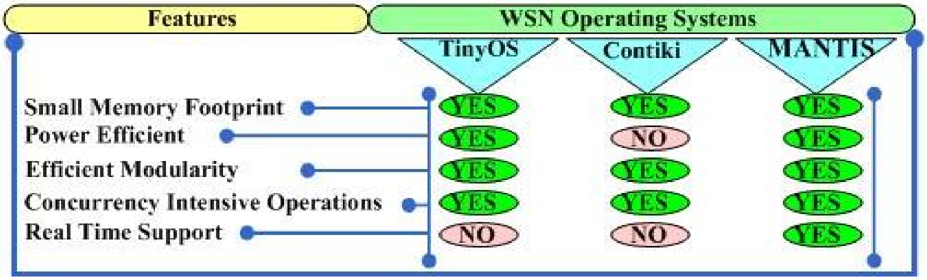
5.4 Memory Management and Security

MANTIS allows dynamic memory management but it discourages to do so because dynamic memory management incurs lot of overhead. Since, memory is a scarce resource in senor nodes therefore; MANTIS OS discourages the dynamic memory management mechanisms. MANTIS manages different threads memory using the thread table that has already been discussed. MANTIS does not provide any mechanism for memory security.

5.5 Communication Protocol Support

MOS implements network stack in two parts. The first part of the network protocol stack is implemented in user space as shown in Figure 4. First part contains the implementation of layer 3, layer 2 and layer 1 protocols. While second part contains the implementation of the MAC and PHY layer operations. The rational behind implementing the layer 3 and above layers functionality in user space is to provide flexibility. If an application wants to use its own data driven routing protocol, then it can implement its routing protocol in the user space and can check its functionality. The downside of the approach is performance i.e., network protocol stack has to use API's provided by MANTIS instead of communication directly with the device driver and hardware. This results in many context switches that involves computational and memory overheads.

The second part of the networking protocol stack is implemented in a COMM layer. COMM layer primarily implements synchronization and MAC layer functionalities. COMM layer provides a unified interface for communication device drivers, for interfaces such as serial, USB, and radio devices. The COMM layer also performs packet buffering functionality. It is possible that packets arrive from the network for a thread that is not currently scheduled. In such scenarios COMM layer will buffer packets. Once the thread gets scheduled COMM layer passes a pointer to the data to the concerned thread.



Features	WSN Operating Systems		
	TinyOS	Contiki	MANTIS
Small Memory Footprint	YES	YES	YES
Power Efficient	YES	NO	YES
Efficient Modularity	YES	YES	YES
Concurrency Intensive Operations	YES	YES	YES
Real Time Support	NO	NO	YES

Fig. 5. WSN OS Grading

6 Future Research Directions

Plenty of research has been done on WSN OS but still it's not an out dated research domain. It's relatively new research domain therefore; there are some issues that need to be resolved. Following are the some issues that must be taken up for future research.

6.1 Support for Real Time Applications

There are many real time application areas for WSN e.g., in industry automation, chemical processes, and multimedia data processing and transmission. Schedulers have been designed to support soft real time operations in some operating systems but the effort is far from complete. In future, we need scheduling algorithms that can accommodate both soft and hard real time requirements of applications.

6.2 Secondary Storage Management

With the passage of time new application areas for WSN are emerging and applications are requiring more and more memory. Typical databases application requires a secondary storage with sensor nodes. According to the best of authors knowledge, there exist no work on secondary storage and file management in sensor nodes. Secondary storage management can be an active area of research for WSN OS in future.

6.3 Virtual Memory

Since, sensor node has very limited RAM and applications are requiring more and more RAM. Therefore, in future we need to introduce a virtual memory concept in sensor networks OSs. We need to device virtual memory management techniques that are power as well as memory efficient.

6.4 Memory Management and Security

Little work has been done on memory management in WSN OS. The primary reason behind this is that, it has been assumed that only single application runs on a WSN OS. In future, we can have sensor nodes that can sense different phenomenon's therefore, it is possible that multiple application runs on sensor node. In such a scenario we need to manage node's memory and we need to protect one process memory from another. Research needs to be done in memory management and security keeping in view the limitations of the sensor nodes.

7 Conclusions and Future Work

In this paper, we have investigated the most widely used operating systems for WSNs. This paper helps to understand the characteristics of an OS for WSNs in particular and embedded devices in general. Design strategies for various components of an OS for WSN has been explained, investigated along with their relative pros and cons. Target application areas of different WSN OS has been pointed out. We believe that presented pros and cons of different design strategies presented here will motivate the researcher to design more robust OSs for WSNs. Moreover, this survey will help the application and network designer to select an appropriate OS for their WSN applications.

In future, we plan to investigate other OSs for WSN i.e, SensorOS [6], A Dynamic Operating System for sensor nodes [7], and Nano-RK [9].

References

1. Levis, P., Madden, S., Polastre, J., Szewczyk, R., Whitehouse, K., Woo, A., Gay, D., Hill, J., Welsh, M., Brewer, E., Culler, D.: Tinyos: An operating system for sensor networks, pp. 115–148 (2005), http://dx.doi.org/10.1007/3-540-27139-2_7
2. Cooprider, N., Archer, W., Eide, E., Gay, D., Regehr, J.: Efficient memory safety for tinyos. In: 5th international conference on Embedded networked sensor systems. SenSys 2007, pp. 205–218. ACM, New York (2007)
3. Klues, K., Liang, C.J.M., Paek, J., Musaloiu, R., Levis, P., Terzis, A., Govindan, R.: TOSThread: Thread-safe and Non-Invasive Preemption in TinyOS. In: 7th ACM conference on Embedded Networked Sensor Systems, pp. 127–140 (2009)
4. Bhatti, S., Carlson, J., Dai, H., Deng, J., Rose, J., Sheth, A., Shucker, B., Gruenwald, C., Torgerson, H.R.: Mantis os: an embedded multithreaded operating system for wireless micro sensor platforms. Mob. Netw. Appl. 10(4), 563–579 (2005)
5. Dunkels, A., Gronvall, B., Voigt, T.: Contiki - a lightweight and flexible operating system for tiny networked sensors. In: 29th Annual IEEE International Conference on Local Computer Networks, pp. 455–462. IEEE Computer Society, Washington (2004)
6. Kuorilehto, M., Alho, T., Hannikainen, M., Hamalainen, T.D.: SensorOS: A New Operating System for Time Critical WSN Applications. In: Vassiliadis, S., Bereković, M., Hämäläinen, T.D. (eds.) SAMOS 2007. LNCS, vol. 4599, pp. 431–442. Springer, Heidelberg (2007)
7. Han, C.C., Kumar, R., Shea, R., Kohler, E., Srivastava, M.: A Dynamic Operating System for Sensor Nodes. In: 3rd International Conference on Mobile systems, applications and services, pp. 163–176 (June 2005)
8. Kim, H., Cha, H.: Multithreading Optimization Techniques for Sensor Network Operating Systems. In: 4th European conference on Wireless Sensor Networks, pp. 293–308 (January 2007)
9. Eswaran, A., Rowe, A., Rajkumar, R.: Nano-RK: an Energy-aware Resource-centric RTOS for Sensor Networks. In: 26th IEEE International Real Time Systems Symposium, pp. 256–265 (December 2005)
10. Reddy, V., Kumar, P., Janakiram, D., Kumar, G.A.: Operating Systems for Wireless Sensor Networks: A Survey. International Journal of Sensor Networks 5(4), 236–255 (2009)
11. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless Sensor Networks: a survey. Computer Networks 38(4), 393–422 (2002)
12. Romer, K., Mattern, F.: The design space of wireless sensor networks. IEEE Wireless Communication 11(6), 54–61 (2004)
13. Hill, J., Szewczyk, R., Woo, A., Hollar, S., Culler, D., Pister, K.S.J.: System architecture directions for networked sensors. In: Architectural Support for Programming Languages and Operating Systems, pp. 93–104 (2000)
14. TinyOS Network Working Group, Web page, http://docs.tinyos.net/index.php/TinyOS_Tutorials#Network_Protocols
15. Network Protocols- TinyOS documentation Wiki. Web page, http://docs.tinyos.net/index.php/Network_Protocols

16. Lin, K., Levis, P.: Data Discovery and Dissemination with DIP. In: 7th International Conference on Information Processing in Sensor Networks, pp. 433–444 (2008)
17. Gay, D., Levis, P., Von Behren, R., Welsh, M., Brewer, E., Culler, D.: The nes C language: A holistic approach to networked embedded systems. In: SIGLAN 2003 (2003)
18. Dwivedi, A.K., Tiwari, M.K., Vyas, O.P.: Operating Systems for Tiny Networked Sensors: A Survey. International Journal of Recent Trends in Engineering 1(2) (May 2009)
19. Protothreads- Lightweight, Stackless Threads in C, `http://www.sics.se/~adam/pt/`
20. Tsiftes, N., Eriksson, J., Dunkels, A.: Low-Power Wireless Ipv6 Routing with ContikiRPL. In: ACM/IEEE IPSN (2010)
21. Winter, T., Thubert, P.: RPL: Ipv6 Routing Protocol for Low Power and Lossy Networks, July 28 (2010) draft-ietf-roll-rpl-11
22. Contiki Documentation, `http://www.sics.se/~adam/contiki/docs/`
23. Von Behren, R., Condit, J., Brewer, E.: Why Events are a Bad Idea (for High Concurrency Servers). In: 9th Workshop on Hot Topic in Operating Systems, HOTOS IX (2003)

A Study on the Implementation of Pigpen Management System Using Wireless Sensor Networks

Jeonghwan Hwang, Jiwoong Lee, Hochul Lee, and Hyun Yoe*

School of Information and Communication Engineering,
Sunchon National University, Korea
{jhwang,leejiwoong,hclee,yhyun}@sunchon.ac.kr

Abstract. The wireless sensor networks (WSN) technology based on low power consumption is one of the important technologies in the realization of ubiquitous society. When the technology would be applied to the agricultural field, it can give big change in the existing agricultural environment such as livestock growth environment, cultivation and harvest of agricultural crops. This research paper proposes the 'Pigpen Management System' using WSN technology, which will establish the ubiquitous agricultural environment and improve the productivity of pig-raising farmers. The proposed system has WSN environmental sensors and CCTV at inside/outside of pigpen. These devices collect the growth-environment related information of pigs, such as luminosity, temperature, humidity and CO2 status. The system collects and monitors the environmental information and video information of pigpen. In addition to the remote-control and monitoring of the pigpen facilities, this system realizes the most optimum pig-raising environment based on the growth environmental data accumulated for a long time.

Keywords: WSN, Ubiquitous, u-IT, Agriculture, Pigpen.

1 Introduction

The wireless sensor networks technology based on low power consumption is one of the important technologies in the realization of the ubiquitous society. It is applied and utilized in various fields such as environment monitoring, disaster control, logistic management and home network [1][2]. As the ubiquitous technology such as wireless sensor network technology has brought big change to other industries and daily living, it can also be utilized in the agriculture and livestock industry in various ways [3].

The ubiquitous agriculture has its purpose on enhancing the productivity by combining IT technology with agriculture, examining the safety of agricultural crops by systematically managing the distribution/consumption of crops and making the process of distribution/consumption transparent [4].

For instance, an unmanned control device is installed in the facility house producing agricultural products. The device automatically measures the environmental change

* Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 632–641, 2010.

factors of the crop, such as temperature, humidity, ammonia gas, CO2 and the weather. The measured information is saved in database and utilized in the agriculture. Through this process, it is possible to reduce the material input for crops growth such as fertilizers and chemicals, decrease the production cost by making the most optimum growth environment and enhance the productivity [5][6].

The ubiquitous technology is also effectively used in the livestock-raising such as pig-raising or chicken-raising. For instance, RFID technology and WSN technology are used in the pig-raising to manage the feeding of pig-individual, pigsty environment and pigs growth tracking. Mobile device gives alarm when there is an abnormality in the pig-individual so that farmers can take immediate action [6].

Recently, domestic pig-raising industry faces a face-to-face duel with pig-raising advanced countries because of rise in the feed cost and the execution of FTA [7]. Also, the mortality rate caused by wasting diseases increases and production cost goes up together, giving double difficulties to pig-raising farmers. Accordingly, the productivity increase and high quality pork production became the essential tasks of pig-raising industry [8].

In order to cope with this issue, it is urgently required to secure the scientific and systematic pig-raising technology combining current u-IT technology with pig-raising industry, which is the primary industry.

This paper proposes ‘Pigpen Management System’ using wireless sensor networks, which is an management system applying the WSN technology to the pig-raising environment management and control.

The proposed system enables the monitoring of pigpen environment and the control of pigpen facilities. Through these, the optimum pig-raising environment can be maintained, productivity can be increased and producer convenience through remote/automatic control can be also achieved.

This research paper is comprised of followings. Chapter 2 will explain the system structure and service process of pigpen management system using wireless sensor network. Chapter 3 will explain the system operation result. Chapter 4 will compare and analyze the experiment result. Finally, Chapter 5 will give conclusion to close the research paper.

2 Design of Pigpen Management System

The Pigpen Management System collects pigpen environment information and video information through environmental sensors measuring environmental elements and CCTV. It also supports the monitoring and control of pigpen status.

2.1 System Architecture

The proposed pigpen management system is comprised of physical layer, middle layer and application layer. The physical layer is comprised of sensors, CCTV and pigpen facilities as in Figure 1. The middle layer supports the communication between physical layer and application layer. It makes the pigpen information into database and provides with monitoring and control service, maintaining the growth environment of pigs at optimum status. The application layer is comprised of interfaces which support the pigpen environment monitoring and pigpen facilities control service.

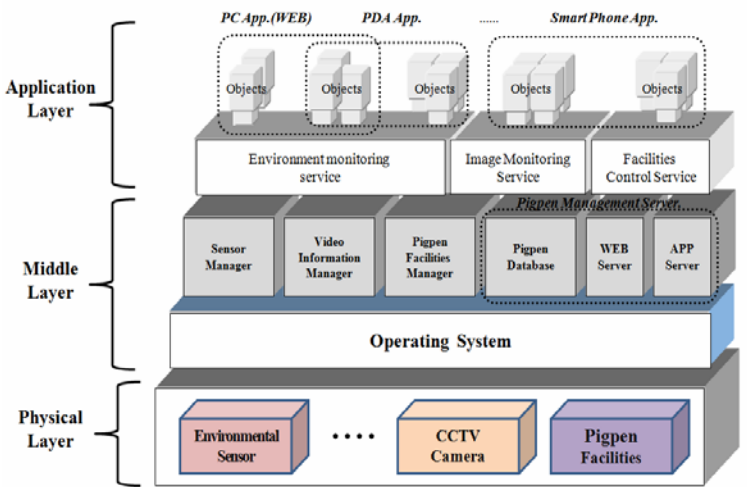


Fig. 1. Pigpen management system architecture

Physical Layer. The physical layer is comprised of environmental sensors, collecting the pigpen environment information, CCTV, collecting the video information of pigpen and pigs, and pigpen facilities, making the optimum growth environment for pigs.

The environmental sensors can be classified into pigpen internal sensors and pigpen external sensors. Internal sensors measure the pigpen internal environmental information such as luminosity, temperature, humidity and CO2. External sensors measure the external environment change of pigpen.

CCTV in the pigpen collects the video information of pigpen and pigs. Pigpen facilities are comprised of lightings, humidifier, air conditioner and ventilator which control the pigpen environment that gives impact on the growth of pigs such as luminosity, temperature, humidity and CO2.

Middle Layer. The middle layer is comprised of sensor manager, video information manager, pigpen facilities manager and pigpen management server. The sensor manager manages he environmental information collected at the sensors in the physical layer. Video information collected at CCTV is managed by video information manager. Pigpen management server, with pigpen database comprised of pigpen information, monitors and controls pigpen facilities.

The sensor manager does 'format processing’, which is changing the pigpen environmental information collected at the environmental sensors of physical layer into a format that can be saved in the pigpen database, ‘units change’, which is changing the units to meet with measurement elements, and ‘update query’, which is processing the data to save in the pigpen database.

The pigpen facilities manager receives the control signal and operates/manages pigpen facilities. It also saves the pigpen facilities status in the pigpen database. The video information manager provides web with stream data.

The pigpen database saves ‘pigpen facilities environment data’, collected at the sensors installed inside/outside of pigpen such as luminosity, temperature, humidity and CO2, ‘video data’, collected at CCTV, ‘pigpen facilities status/control data’ and ‘environmental standard values’ for the automatic control and status notification, in the tables allocated to each of them.

The pigpen management server is located between the producer and pigpen database. It examines the environmental data saved in the pigpen database in fixed cycle, reports them to the producer and controls the pigpen facilities by comparing them with the environmental standard values saved in the pigpen facilities control table.

Application Layer. The application layer is comprised of application services supporting various platforms such as laptop, web, PDA and smart phones. It provides producer with 'pigpen environment monitoring service', 'pigpen video monitoring service' and 'pigpen facilities control service'.

2.2 Services of Pigpen Management System

This system provides with ‘pigpen environment monitoring service’, enabling the observation of internal/external environmental information of pigpen, ‘pigpen video monitoring service’, providing with pigpen video in real time, ‘pigpen facilities control service’, enabling the automatic control and manual control of pigpen facilities by producer based on the environmental standard values, and ‘danger alarm service’, giving notification of dangerous situation at the pigpen.

Pigpen Environment Monitoring Service. The pigpen environment monitoring service shows the pigpen environmental data, collected at the environmental sensors measuring the environmental elements, such as luminosity, temperature, humidity and CO2, to producer through GUI so that producers can identify the environment changes of internal and external of the pigpen.

The detail of this service is that it collects pigpen internal/external environmental information giving impacts to pigs growth such as luminosity, temperature, humidity and CO2 from the environmental sensors installed at inside/outside of pigpen and transmits the information to sensor manager periodically.

The sensor manager will analyze the received data and extract each sensing value. Their formats will be changed and they will be saved in each table of pigpen database. The pigpen management server transmits pigpen internal/external environmental information saved in the pigpen database to producer and the producer can monitor the environmental information of pigpen through this information.

Pigpen Video Monitoring Service. The pigpen video monitoring service provides producer/consumer with video of pigpen/pig-individuals through CCTV installed in the pigpen.

The CCTV sends the pigpen video to video information manager and the video information manager provides with this information by web through Internet. Users can confirm the pigpen video information through Internet.

Pigpen Facilities Control Service. The pigpen facilities control service enables the pigpen management server automatically control the pigpen facilities, or, the producer manually control the pigpen facilities based on the collected information at the CCTV and environmental sensors installed at inside/outside of pigpen.

The automatic control service saves the information collected from pigpen at pigpen database. The pigpen management server calls up the information and compares it with the environmental standard values saved in the pigpen database. If it is more than or short of standard value, it will confirm whether the pigpen facilities are operating as saved in the pigpen database. Then it will send the control signal to pigpen facilities manager and control the pigpen facilities.

When pigpen facilities operate, the pigpen facilities status information is saved in the pigpen database and it will be notified to user.

The manual control service saves the information collected from pigpen in the pigpen database and the pigpen management server sends the information to the user in real time.

If the user wants to control the pigpen at this time, the user will send the pigpen facilities control signal to pigpen management server through GUI. The pigpen management server will check whether the pigpen facilities are operating through pigpen database and send the control signal to pigpen facilities manager to control the pigpen facilities.

Danger Alarm Service. The danger alarm service tells the weather change and pigpen status change to farmers in real time and takes emergency measure to prevent danger in advance. The data sensed at the environmental sensor is sent to the sensor manager. The sensor manager extracts the sensing values from received data and saves them in the pigpen database. The saved sensing values will be periodically monitored by pigpen management server. If it would be more than or less than the standard value, it will be notified to the element where the event had occurred.

3 Implementation of Pigpen Management System

3.1 Components of Pigpen Management System

Environmental Sensors. In order to collect the environmental information of pigpen, WSN environmental sensors were installed at inside/outside of pigpen. These sensors will form the wireless network together with WSN sensor gateway in the pigpen. The sensors are classified into ‘integrated sensor node’, measuring the temperature, humidity and luminosity, and 'CO2 nodes', measuring CO2.

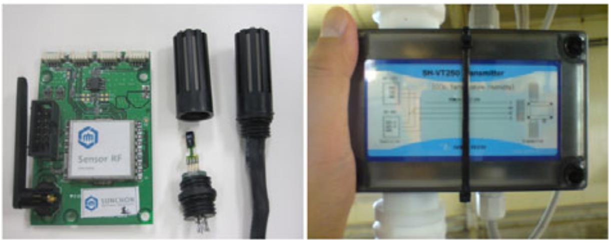


Fig. 2. Integrated sensor node and CO2 sensor

The integrated sensor node receives the sensor data from temperature, humidity sensors. It processes the data at MSP430 MCU and transmits them to relay node and gateway, using CC2420 RF chip. In order to reduce the heat impact the sensor receives from the node, the node and the sensor will maintain certain distance from each other.

MSP430 is 16bit RISC with 48Kbyte 'program memory' and 10Kbyte RAM inside. It can process multiple sensor data at high speed. CC2420 is RF chip supporting Zigbee. It supports the frequency band of 2400~2483.5 MHz. It operates in DDDS method, supports O-QPSK modulation method and 250k bps baud rate. It enables real time wireless communication with small power consumption.

SHT71 was used for temperature/humidity sensors. SHT 71 temperature/humidity sensor has temperature sensor and humidity sensor in one body. It works on relatively small power source of 2.4V~5.5V and has small power consumption of average 28 uA. Having correction memory, it has 14bit A/D converter and digital 2-wire interface. It measures temperature from '-40°' to '120°' with 0.5° error accuracy. Humidity can be measured between 0% and 100% with 3.5% error accuracy.

3.3V operating voltage was connected to integrated sensors node and digital 2-wire was connected to MSP430 circuit to process the temperature and humidity information of pigpen.

CO2 sensor uses NDIR measurement method. It measures the range of 0 to 3,000 ppm with 3% error accuracy. RS485 method was used for communication method.

CCTV. In order to monitor the pigpen by 24 hours video, monitoring camera based on IP was installed as in Figure 3. This camera can monitor the pigpen status in real time and is also used to find out the cause of accident, in case there was an accident such as theft or accident in the pigpen, by monitoring and recording the pigpen inside 24 hours. The recorded video information is transmitted to the 'pigpen management server', where they are saved in the database after classification by pigpen ID and camera number.

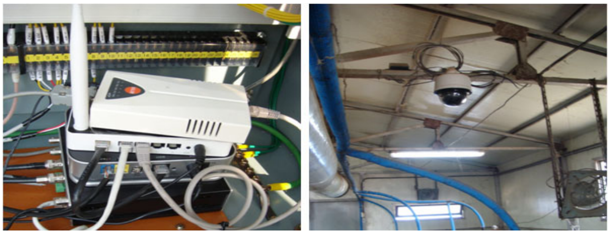


Fig. 3. DVR and CCTV camera

Pigpen Facilities and Environmental Control Device. Luminosity, temperature, humidity and CO2 give impacts on the growth of pigs. Figure 4 shows the environmental control devices in the pigpen, which enables the control of pigpen facilities such as lighting, humidifier, fan heater, air conditioner and ventilator for those. Through these environmental control devices, it is possible to maintain pleasant pig-raising environment in the pigpen.

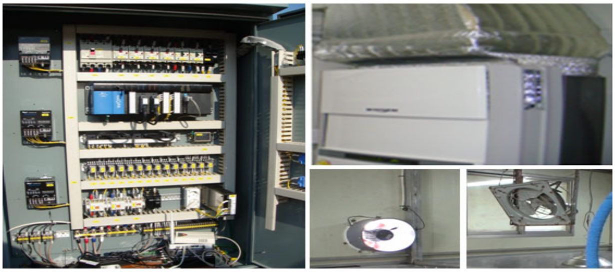


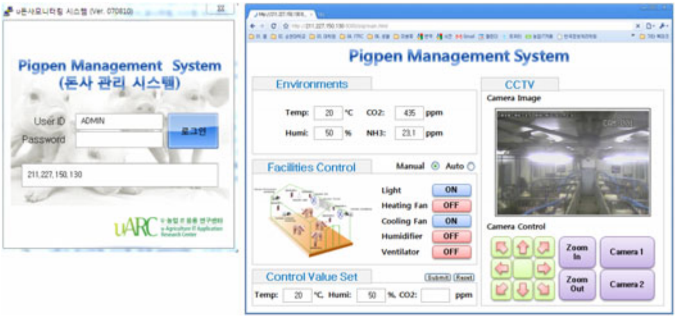
Fig. 4. Environmental control device and Pigpen facilities

Pigpen Management Server. The environment measurement data in stream-form transmitted from pigpen is parsed and saved in the database. At the same time, it is sent to the manager in charge of relevant pigpen so that he/she can know the environment change in real time.

The process of pigpen environmental status of pigpen is classified into two processes. First one is automatically controlling the environmental status from system in reference to the designated environmental standard data. Second is directly controlling the system dependent on the necessity of the manager.

Applications. GUI for manager is developed for web environment. Tomcat-6.0.20 is used for WAS and 'mysql' is used for database. The latest released version 5.0 was used. It was possible to collect the pigpen environmental information and video information of pigpen through sensors and video monitoring camera and constantly monitor/control the pigpen status through user-intuitive GUI by way of above result.

Figure 5 shows the Web GUI of pigpen management system.



Pigpen Management System
(돈사 관리 시스템)

User ID ADMIN

Password

211.227.150.130

Pigpen Management System

Environments

Temp: 20 °C CO2: 435 ppm

Humi: 50 % NH3: 23.1 ppm

Facilities Control Manual ● Auto ○

Light ON

Heating Fan OFF

Cooling Fan ON

Humidifier OFF

Ventilator OFF

Control Value Set

Temp: 20 °C, Humi: 50 %, CO2: ppm

Fig. 5. Web Graphic User Interface of Pigpen Management System

4 Results

4.1 Measurement Environment

In order to complete the design and evaluate the performance of this system, two test-beds were established in two pigsties. Pigpen A had sensors and WSN. Pigpen B had sensors, WSN and ‘pigpen management system’. Both pigpens were located in the same area with 5m distance from each other.

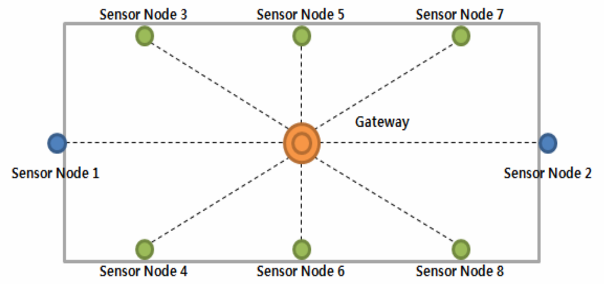


Fig. 6. Wireless sensor networks topology

Figure 6 is the structure of sensors and gateways installed in the two pigpens. Sensor 1 and sensor 2 are located outside of pigpen to measure the external temperature and humidity. Sensors 4 to 6 were installed inside pigpen to measure the internal temperature and humidity.

4.2 Measurement Results and Analysis

The environment sensors installed at pigpen send measured data to server every 10 minutes. The server applies 'the pigpen environment decision making method' to the measured data and control the pigpen environment. Measurement period was from 00:00 hours of April 1st, 2010 until 23:00 hours of April 3rd. The measured environmental data are shown in graph with one hour interval.

Figure 7 is the variations of temperature and humidity in pigpen A using existing control method.

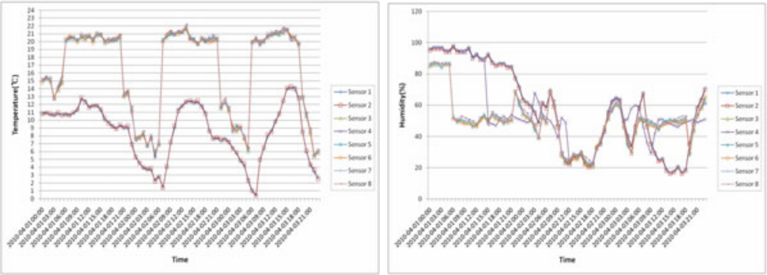


Fig. 7. Variations of temperature and humidity in pigpen A

The measurement result of pigpen A suggests that the temperature and humidity of pigpen inside keep constant level from 06:00 hour to 20:00 hour, because the producer directly controls the pigpen facilities. However, from 20:00 hour to 05:00 hour, there are rapid changes in the temperature and humidity caused by absence of proper pigpen facilities control. Such rapid environmental change in the pigpen gives severe stress to pigs, which can lead to the deaths of pigs.

Figure 8 is the variations of temperature and humidity in pigpen B which has operating ‘pigpen management system’.

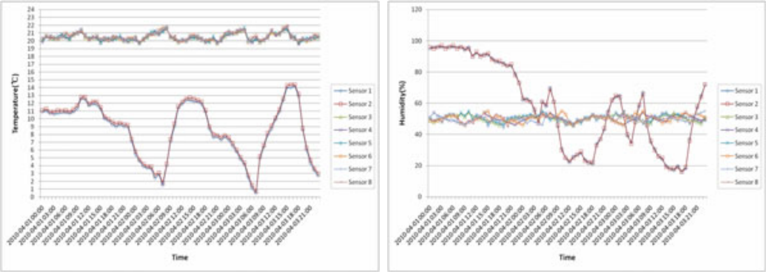


Fig. 8. Variations of temperature and humidity in pigpen B

Pigpen B generates the estimated data based on the measured data by way of ‘pigpen management system’ proposed by this research. It activates the pigpen internal control devices and it was possible to maintain the pigpen internal temperature and humidity near to the environmental standard values. Pigpen B showed uniform temperature and humidity status compared to pigpen A using existing control method.

This measurement result suggests that the pigpen operation by proposed pigpen management system is more effective than the pigpen operation by existing control method.

5 Conclusions

This research proposed ‘Pigpen Management System’ using wireless sensor networks as the system to manage the pigpen environment in integrated way in the ubiquitous agricultural environment.

The pigpen management system is comprised of three layers. The roles and services provided by these three layers have been explained. The comprising elements of each layer provide user with organic information by collecting and managing the environmental elements in the pigpen. The system provides with management devices proper to pigpen, saves the provided data, makes a 'control manual' and keeps the record, so that there would not be trial-and-error situation even if the person in-charge would be replaced. It is expected that the mortality rate of pigs could be reduced substantially.

The pigpen management system using wireless sensor networks will contribute in the saving of labor force in the pig-raising farmers, production of high quality pork and further contribute in the securing competitiveness of pig-raising industry, by way of joining pig-raising industry with ubiquitous technology.

Acknowledgements. This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2010-(C1090-1021-0009)).

References

1. Akyildiz, I.F., et al.: A survey on Sensor Networks. IEEE Communications Magazine 40(8) (2002)
2. Chong, C.-Y., Kumar, S.P., Hamilton, B.A.: Sensor networks: evolution, opportunities, and challenges. Proc. IEEE 91(8), 1247–1256 (2003)
3. Pyo, C.-S., Chea, J.-S.: Next-generation RFID / USN technology development prospects. Korea Information and Communication Society, Information and communication, 7–13 (2007)
4. Shin, Y.-S.: A Study on Informatization Model for Agriculture in Ubiquitous Era, MKE Research Report (2006)
5. Lee, M.-H., Shin, C.-S., Jo, Y.-Y., Yoe, H.: Implementation of Green House Integrated Management System in Ubiquitous Agricultural Environments. Journal of KIISE 27(6), 21–26 (2009)
6. Yoo, N., Song, G., Yoo, J., Yang, S., Son, C., Koh, J., Kim, W.: Design and Implementation of the Management System of Cultivation and Tracking for Agricultural Products using USN. Journal of KIISE 15(9), 617–674 (2009)
7. Lee, J.-H.: [Livestock industry research series 11] What is a threat to South Korea's livestock industry? Focus attention GSnJ No.55 (2008)
8. Yoo, Y.-H., Kim, D.-H.: The current state of automation in pig house establishment and prospection. Korea society for livestock housing and environment 19, 29–47 (2006)

A Study on Energy Efficient MAC Protocol of Wireless Sensor Network for Ubiquitous Agriculture

Ho-chul Lee, Ji-woong Lee, Jeong-hwan Hwang, and Hyun Yoe*

School of Information and Communication Engineering,
Sunchon National University, Korea
{hclee,leejiwoong,jhwang,yhyun}@sunchon.ac.kr

Abstract. Various technologies are used in the agricultural sites now. Especially, the recent application of sensor network related technology is quite notable. Considering the efficiency of MAC protocol of WSN is being researched in various aspects, it is believed that a research on how to apply the MAC protocol to agriculture would be also required. This research is based on the sensor node developed by Sunchon University ITRC. Once the sensor nodes are effectively located in the farm, they operate for a long time and they are rarely relocated once installed. The concentration of multiple sensor nodes in a narrow area is another characteristic the sensor node. The purpose of this research is to select a sensor network MAC protocol, which would be most proper to agricultural site with good energy efficiency and excellent transmission delay performance. The applicable protocols such as S-MAC and X-MAC were set up for the installation environment. They were compared and a methodology to select the most optimum protocol to agricultural site is suggested.

Keywords: WSN, Ubiquitous, u-IT, cultivation facility, Paprika.

1 Introduction

The recent innovation in IT technology is accelerating the fusion between industries. The fusion between IT and traditional industries continuously goes on. The application of ubiquitous technology to agriculture, which is a primary industry, is getting expectation that the convergence technology would enhance the added-value and productivity of agriculture [1]. In order to establish such u-agriculture environment successfully, the core ubiquitous technology development optimized to agriculture, such as sensor hardware, middleware platform, routing protocol and agriculture environment application service, would be essentially required [2].

For the development of such ubiquitous technology, various energy-efficient MAC protocols were studied in the wireless sensor network. S-MAC[3], applying “sleeping, stand-by”, was suggested to improve the energy efficiency of MAC protocol. T-MAC[4] was suggested to reduce the unnecessary waking hours even a little bit more.

* Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 642–650, 2010.

Adaptive S-MAC [5] was developed to avoid the transmission delay phenomenon occurring when applying the duty cycle and hybrid type Z-MAC [6] was developed combining CSMA and TDMA. There is also the X-MAC [7], which preoccupies the channel using preamble during the sleeping period in asynchronous method.

This research chooses the MAC protocol, which demonstrates the most efficient energy performance when WSN would be applied to paprika cultivation in a cultivation facility. Further, a methodology to choose MAC protocol proper to certain cultivation method or stock raising method will be suggested. Actual cultivation facility was taken as the model for the research and sensors were located proper to the cultivated crop. The network topology of the sensors was configured and sensors performance will be measured by a simulator.

Paprika cultivation facility was chosen because paprika takes an important role among Korean major exporting horticultural products. Paprika is being exported to Japan and United States, Russia and Taiwan are potential export markets [8]. Paprika is a tropical garden fruit. The harvest quantity of paprika shows big variation, dependent on sunlight, temperature and humidity environments in the cultivation facility, in addition to cultivation and management technology [9]. Especially, the harvest-cycle variation range of paprika is very big, dependent on the number of fruit-setting caused by interaction between luminosity and temperature [10]; therefore, very precise control of luminosity and temperature is required. When the productivities of paprika in the plastic film house and glass greenhouse were compared, the glass greenhouse showed twice productivity of plastic film house [11]. The productivity of paprika in Korea were 6.8kg/㎡ in year 2000 and 9.4kg/㎡ in year 2007. Even there was 38% productivity increase during seven years; it is still 30% of average productivity in Netherlands. [12].

This research paper comprised of followings. The MAC protocols to be compared and analyzed in this research will be introduced in Chapter 2. In Chapter 3, the methodology of selecting MAC protocol to be used at cultivation site is suggested and candidate MAC protocols are compared. Then there will be a conclusion section.

2 Relate Works

2.1 S-MAC

SMAC is a representative synchronous MAC protocol. It periodically repeats inactivated “sleep mode” and activated “listen mode” with fixed lengths [3]. In the “listen mode”, the data transmission between two nodes is possible. In the “sleep mode”, power waste at each sensor node is reduced by providing with minimum power to maintain the sensor node, while main power is shut-off. However, there will be “listen-sections” without communication caused by fixed lengths and power is wasted because of these unused “listen-sections”. Also, there is a disadvantage in the “sleep-section”, which is data transmission delay caused by inability to receive signal in the “sleep-section”.

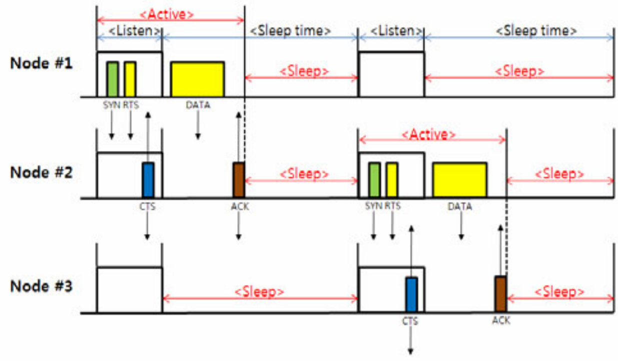


Fig. 1. S-MAC

2.2 X-MAC

The X-MAC protocol is suggested to resolve the problem of overhearing caused by the long preamble used in the B-MAC [13] protocol. It reduces the preamble overhearing of B-MAC protocol by repeated transmission of minimum preamble for synchronization and the "strobed preamble" containing the destination address. When there is data to transmit, the node operating with X-MAC transmits the minimum preamble and the "short preamble", containing the destination address, in order to tell nearby nodes that it has data to transmit. Then the node maintains "stand-by" mode of data reception for a sufficient period to receive early ACK [7].

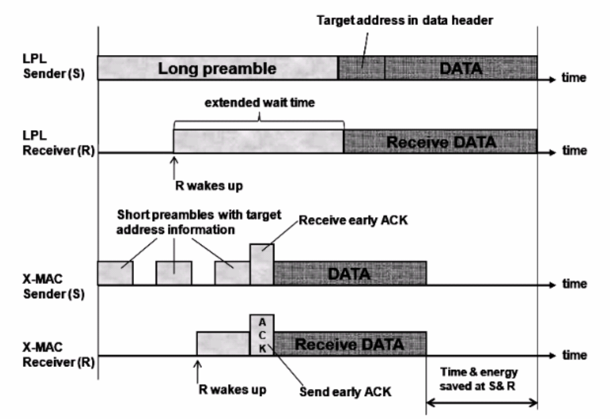


Fig. 2. Comparison of the timelines between LPL's extended preamble and X-MAC's short preamble approach[7]

3 Method of Protocol Adaptation

3.1 Cultivation Environment of Paprika

This research takes the actual paprika greenhouse in Gwangyang, Chollanam-do. Sensor nodes are located in the greenhouse and the network performance was examined in advance so that the sensor network can be installed for efficient operation by choosing the efficient MAC at site.

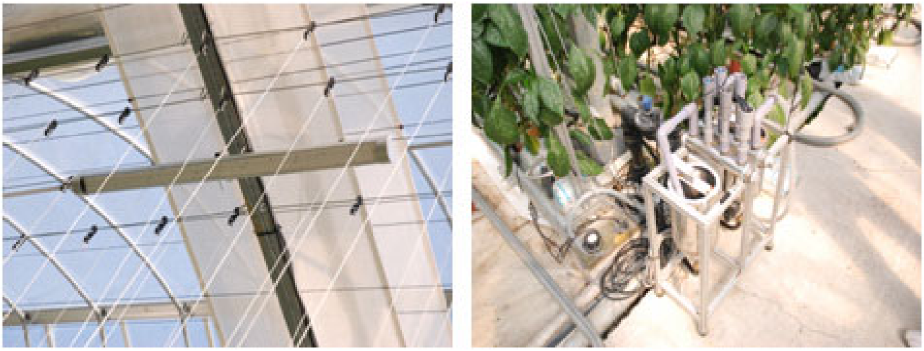


Fig. 3. LED Lamp, Sun Shield and Warm water supply

In the facility cultivation, paprika seeds are sown to the rock wool trays and they are planted temporarily on the rock wool cubes. When there would be the first branch stem, they will be permanently planted to the culture medium for cultivation. Then culture solution made for the best paprika growth will be supplied. At the lower part of the culture medium, boiler pipe way will be installed and warm water will be supplied to maintain the temperature at the paprika rooting zone constant. LED lightings will be installed in the upper part to enhance the growth of paprika and ventilation equipment will be also installed to mix the upper air and lower air for constant temperature. Mobile screens will be installed inside of greenhouse roof to shut off the strong sunlight.

The paprika rock wool cubes will be located at 30cm distance with each other. There will be four rock wool cubes for each culture medium. Each culture medium will make 50m length in parallel with the rail installed at greenhouse floor. Each group of culture media will be located at 50cm distance having the rail between them. Other than moving and working space in the greenhouse, the whole greenhouse will be filled with culture media planted with paprika. For growth environment, daytime 22~25℃, nighttime 18~20℃ and humidity 70~75% will be maintained.

3.2 Hardware Description

The sensor node developed by Sunchon National University ITRC Research Center will be applied to this research. This sensor node can collect the information of leaf

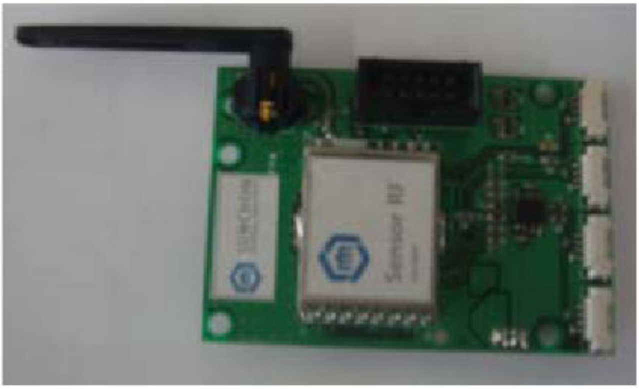


Fig. 4. Sensor Node

wetness, leaf temperature, greenhouse temperature/humidity and control the relay by one sensor. MSP430 MCU is applied to the CPU and CC2420 RF module of Chipcon Co. is used as the data transmission/reception device. The MSP430 microprocessor has 16 bit RISC structure and it works in very fast speed with its 48 Kbyte program memory and 10Kb RAM. 3.6V battery is used for power supply [14].

3.3 Network Topology

The sensor nodes will be installed at every 5m along the culture media lined up in reference to the paprika at the most outer side. They will be installed alternately for the root zone parts and upper parts. Installation will continue to the culture media with 5m distance in reference to the culture media with sensors installed. The overall

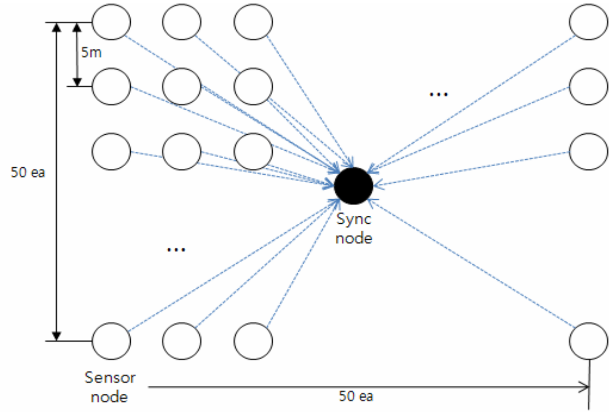


Fig. 5. Network Topology

location shape of sensor nodes is grid-type with 5m distance. The sink node to transmit collected data to the server will be located in the center of 50 * 50 grids. The shape of sensor nodes location is grid shape; however, the network topology is a star topology in reference to the sink node in the center of the grid.

3.4 Duty Cycle

The sensor of sensor node measures leaf temperature, leaf wetness and greenhouse temperature/humidity. They are measured in 3 minutes cycle and transmitted to the server. The relay control port of the sensor node will not be used. The measurement cycle can be different dependent on the characteristic of the crops.

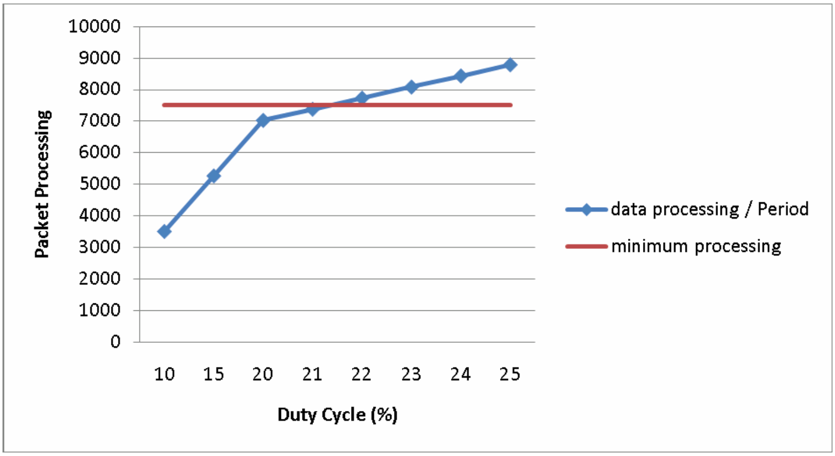


Fig. 6. Available Duty cycle of S-MAC

Now, the duty cycle to be applied to sensor node will be determined. The generated data and the number of nodes to transmit data to sink node, during the measurement cycle of crop environment data, will be estimated. Total number of data which can be generated during the measurement cycle will be estimated and the data quantity which can be processed for each duty cycle will be deduced. The duty cycle will be chosen so that it can process more data than the data generated during the measurement cycle. 10% allowance will be given so that waste data would not be generated. If the measured data would be missed, the on-time responding to the change of crops growth environment change will be difficult and the quality, quantity of the harvested crop will get negative impact.

Figure 6 and Figure 7 shows the packets which can be processed by S-MAC and X-MAC, dependent on duty cycles. In the established situation, the effective duty cycle of S-MAC is 22%, including 10% allowance. For X-MAC, the proper duty cycle is shown as 16.3%.

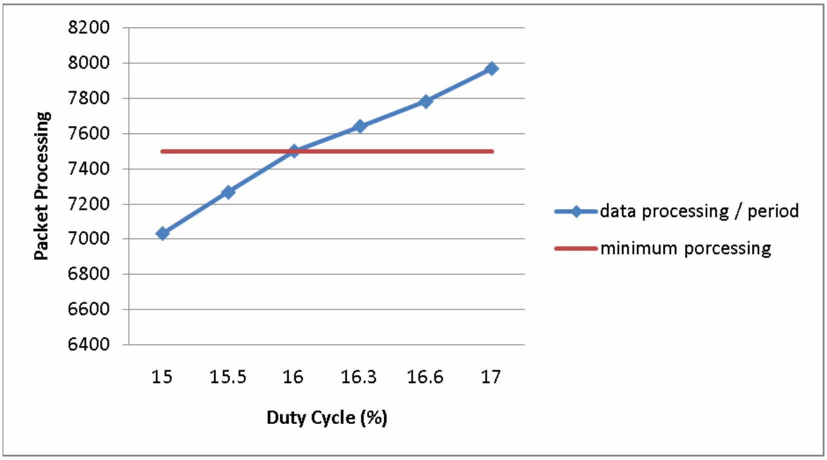


Fig. 7. Available Duty cycle of X-MAC

3.5 Simulation

A simulation will be done to measure the energy performance of the MAC considered for the application. The performances of S-MAC and X-MAC will be examined. For this, a simulation environment was made using NS-2 [15]. [Table 1] is the system parameters for the simulation.

Table 1. Simulation Parameter

	S-MAC	X-MAC
NS2 Version	NS-2.34	
Simulation Time	5000 Second	
Packet Size	40 byte	
Packet Interval	1 minute / node	
Node Count	2500	
Routing Protocol	DSDV	
Duty Cycle	22%	16.3%
Bandwidth	250Kbps	
Initial Energy	30,000 J	

Total 2,500 nodes comprised the network topology as in [Figure 7]. The physical shape of nodes is grid; however, the shape of network topology is a star-shape with sink node in the center. Each node generates 40 bytes per minute of sensing data when the number of the nodes is fixed and the energy consumption at this time is measured. Each node measures leaf temperature, leaf wetness and greenhouse temperature/humidity. They are measured in 3 minutes cycle and transmitted to the server; however, it will be assumed in the simulation that the measurement items will be data with same size and the data is generated in one minute cycle.

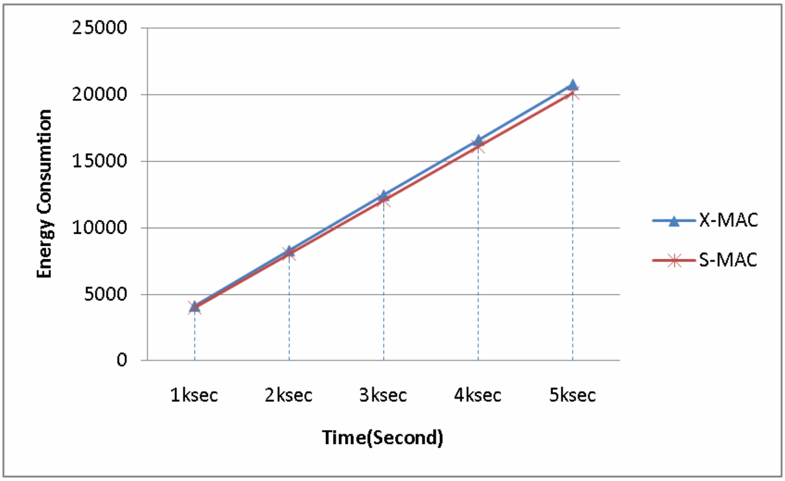


Fig. 8. Energy Consumption

The simulation measurement result in Figure 8 suggested that the energy performance of S-MAC is slightly better than X-MAC. It is believed that this phenomenon occurs when there are many nodes and the number of packets being transmitted is very small. As we can learn from the simulation result, when we apply sensor network to agricultural environment, the protocol proper to the operation environment can be chosen through the comparison and analysis of MAC protocol to be applied, together with the design of network topology in advance.

4 Conclusion

When applying a wireless sensor network to agriculture, the first thing to determine is whether agricultural site is dynamic or static. Next will be the sensors proper to the environment information of the crops to be measured, number of sensors for each sensor node and the data measurement cycle. If one sensor node will have multiple sensors, there will be more power consumption to operate sensor, in addition to the data transmission/reception. The capacity of the sensor to be used would be chosen considering all these.

Then the sensors will be located at the position of the crop to be measured. The location of sink node will be determined by the methodology of data collection. Then the data generation cycle of sensor and the shape of network topology will be determined. After that, the MAC protocol will be determined, applying the MAC protocol determining methodology suggested by this research. It was found that the S-MAC protocol is proper as the MAC to be applied to the facility cultivation of paprika.

This research result suggested the methodology to deduce the most efficient protocol which can be applied to the facility cultivation of paprika. However, it is believed that this result can be also applied to the outdoor cultivation, outdoor stock-raising and cultivation of other crops than paprika in the determination of proper MAC protocol for the situation and subsequent efficient operation.

Acknowledgements. This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2010-(C1090-1021-0009)).

References

1. Lee, K.H., Ahn, C.M., Park, G.M.: Characteristics of the Convergence among Traditional Industries and IT Industry. Electronic Communications Trend Analysis 23(2), 13–22 (2008)
2. Lee, M.-h., Shin, C.-s., Jo, Y.-y., Yoe, H.: Implementation of Green House Integrated Management System in Ubiquitous Agricultural Environments. Journal of KIISE 27(6), 21–26 (2009)
3. Ye, W., Heidemann, J., Estrin, D.: An Energy-Efficient MAC Protocol for Wireless Sensor Networks. In: 21st Conf. of the IEEE computer and communicaions Societies (INFOCOM), vol. 3, pp. 1567–1576 (2002)
4. van Dam, T., Langendoen, K.: An adaptive energy efficient mac protocol for wireless sensor networks. In: 1st ACM Conference on Embedded Networked Sensor Systems (SenSys), pp. 171–180 (2003)
5. Ye, W., Heidemann, J., Estrin, D.: Medium access controlwith coordinated, adaptive sleeping for wireless sensor networks. ACM Transactions on Networking 12(3), 493–506 (2004)
6. Rhee, I., Warrier, A., Aia, M., Min, J.: ZMAC: a Hybrid MAC for Wireless Sensor Networks. In: Proc. of 3rd ACM Conference on Embedded Networked Sensor Systems, SenSys 2005 (2005)
7. Buettner, M., Yee, G.V., Anderson, E., Han, R.: X-MAC: A Short Preamble MAC Protocol For Duty-Cycled Wireless Sensor Networks. In: Conference on Embedded Networked Sensor Systems, pp. 308–320 (2006)
8. Korea Agricultural Trade Information(KATI), The state of sweet pepper industry in korea, Korea Agro-Fisheries Trade Corporation (2009)
9. Dorais, M.: The use of supplemental lighting for vegetable crop production: Light intensity, crop response, nutrition, crop management, cultural practices. In: Canadial Greenhouse Conference (2003)
10. Heuvelink, E., Marcelis, L.F.M., Korner, O.: How to reduce yield fluctuations in sweet pepper. Acta. Hort. 633, 349–355 (2004)
11. Jeong, W.-J., Lee, J.H., Kim, H.C., Bae, J.H.: Dry Matter Production, Distribution and Yield of Sweet Pepper Grown under Glasshouse and Plastic Greenhouse in Korea. Journal of Bio-Environment Control 18(3), 258–265 (2009)
12. Jeong, W.J., Kang, I.K., Lee, J.Y., Park, S.H., Kim, H.S., Myoung, D.J., Kim, G.T., Lee, J.H.: Study of dry and bio-mass of sweet pepper fruit and yield between glasshouse and plastic greenhouse. The Kor Soc. Bio-Environ. Control 17(2), 541–544 (2009)
13. Polastre, J., Hill, J., Culler, D.: Versatile low power media access for wireless sensor networks. In: The Second ACM Conference on Embedded Networked Sensor Systems (SenSys), pp. 95–107 (2004)
14. Park, D.-H., Kang, B.-J., Cho, K.-R., Sin, C.-S., Cho, S.-E., Park, J.-W., Yang, W.-M.: A Study on Greenhouse Automatic Control System Based on Wireless Sensor Network. Wireless Pers Commun (2009)
15. Network Simulator, `http://www.isi.edu/nsnam/ns`

Design and Implementation of Wireless Sensor Networks Based Paprika Green House System

Ji-woong Lee, Ho-chul Lee, Jeong-hwan Hwang, Yongyun Cho,
Changsun Shin, and Hyun Yoe*

School of Information and Communication Engineering, Sunchon National University, Korea
{leejiwoong,hclee,jhwang,yycho,csshin,yhyun}@sunchon.ac.kr

Abstract. This research paper suggests the 'Paprika green house system' (PGHS), which collects paprika growth information and greenhouse information to control the paprika growth at optimum condition. The temperature variation range of domestic paprika cultivation facilities are relatively quite big and the facility internal is kept at relatively dry condition. In addition, the concentration of CO2 is not uniform, giving bad impact on the growth of paprika. In order to cope with these issues, the 'Paprika green house system' (PGHS) based on wireless technology was designed and implemented for the paprika cultivating farmers. The system provides with the 'growth environment monitoring service', which is monitoring the paprika growth environment data using sensors measuring temperature, humidity, illuminance, leaf wetness and fruit condition, the 'artificial light-source control service', which is installed to improve the energy efficiency inside greenhouse, and 'growth environment control service', controlling the greenhouse by analyzing and processing of collected data.

Keywords: USN, Paprika, Green house.

1 Introduction

Recently domestic horticultural industry achieved substantial growth both in quantity and quality with its technology and capital-intensive industry characteristic. Now it became a competitive industry with big potential in overseas export demand, in addition to existing domestic demand [1].

Paprika is one of horticultural products that create high added-value. The production quantity of paprika varies dependent on sunlight quantity, illuminance and sunlight hours [2]. The cultivation cost of paprika is comprised of heating cost, agricultural material cost and labor cost. Among them, the weights of heating cost and agricultural material cost are very high, giving difficulty to the cultivating farmers [3].

This research paper suggests the establishment of a 'Paprika green house system' (PGHS) in the paprika-cultivating green houses, which need precise growth management. 'Paprika green house system' (PGHS) utilizes IT technology in collecting the crops-growth environmental-information in real time and controls the environmental system in the cultivation facility. 'Paprika green house system' (PGHS) reduces the deviations in growth, development, production-quantity and quality of crops. It also

[*] Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 651–660, 2010.

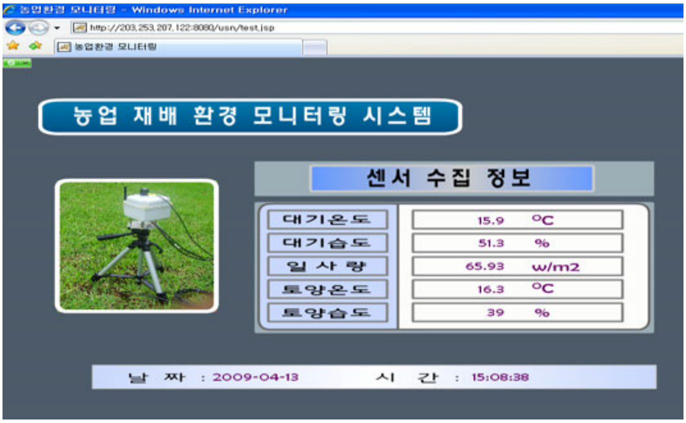
maintains optimum environment in the cultivation facility using biometric data and creates optimum condition at paprika root zone. The system optimizes the management of production elements and reduces the loss of energy, fertilizer and water, which will result in the decrease of production cost. The artificial light-source from artificial lighting makes pleasant growth-environment so that continuous supply of high quality, fresh vegetable would be possible. Farmers will be able to increase the productivity and income by having their cultivation facilities as continuous supply source of high quality fresh vegetable to clients. 'Paprika green house system' (PGHS) is designed and realized to enable all above based on wireless sensor network.

This research paper is comprised of followings. Chapter 2 introduces the technologies related to the monitoring system applied to the agricultural environment in Korea and overseas. Chapter 3 explains the configuration elements and services provision by 'Paprika green house system' (PGHS) suggested by this research. Chapter 4 explains the implementation content of 'Paprika green house system' (PGHS). Chapter 5 will be the conclusion.

2 Related Researches

2.1 Agricultural Monitoring System Using Integrated Sensor Module

This system uses various environmental sensors to collect information required for the cultivation environment of crops. It is a real-time agricultural environment monitoring system based on sensor network. Most of existing wireless sensor nodes based on sensor network need separate conversion/control module for each sensor characteristic. To overcome this issue, an integrated sensor module was developed, which can integrate various sensors used in getting the information required for the crops cultivation, into a single node. New sensors and network monitoring system were also developed to let them fit with the new integrated sensor module and they were integrated to the test environment, in order to examine the operation of newly developed system [4]. Sensor node is also installed to measure the environmental information so that real-time monitoring would be possible [4].


농업환경 모니터링 - Windows Internet Explorer
http://203.253.207.122:8080/usn/test.jsp
농업환경 모니터링

농업 재배 환경 모니터링 시스템

센서 수집 정보

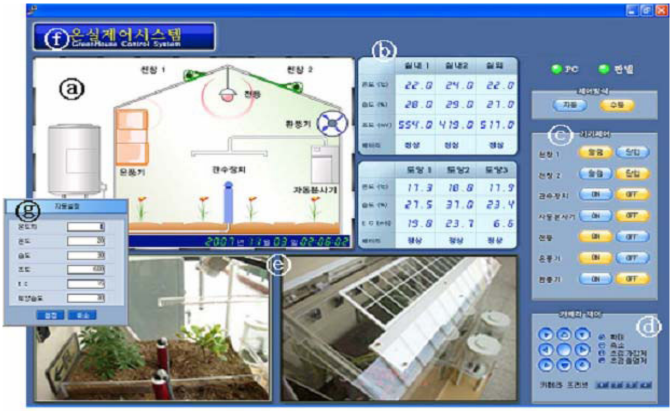
대기온도	15.9 °C
대기습도	51.3 %
일 사 량	65.93 w/m2
토양온도	16.3 °C
토양습도	39 %

날 짜 : 2009-04-13 시 간 : 15:08:38

Fig. 1. Monitoring System

2.2 ‘Greenhouse Environment Integrated Management System’

The ‘greenhouse environment integrated management system’ enables the monitoring of greenhouse status in real time through Internet. With its remote control capability, it enables users to manage their farms without restriction of time and place, as long as Internet connection is available[5]. In order to make ubiquitous agricultural environment, a sensor network was built in the greenhouse to measure the environmental elements affecting cultivation environment, such as temperature, humidity, amount of insolation, CO2, ammonia, wind velocity and precipitation. Also, a ‘greenhouse environment monitoring system’ – comprised of ventilators, windows, heaters, humidifiers, lightings and video processors – is suggested to control the devices activated by the change of measured environmental elements [5].



	실내1	실내2	실외
온도(℃)	22.0	24.0	22.0
습도(%)	28.0	29.0	21.0
조도(mV)	554.0	419.0	511.0
배터리	정상	정상	정상

	토양1	토양2	토양3
온도(℃)	11.3	10.8	11.9
습도(%)	27.5	31.0	23.4
EC(mS)	19.8	23.7	6.6
배터리	정상	정상	정상

Fig. 2. Monitoring System GUI

3 Paprika Green House System

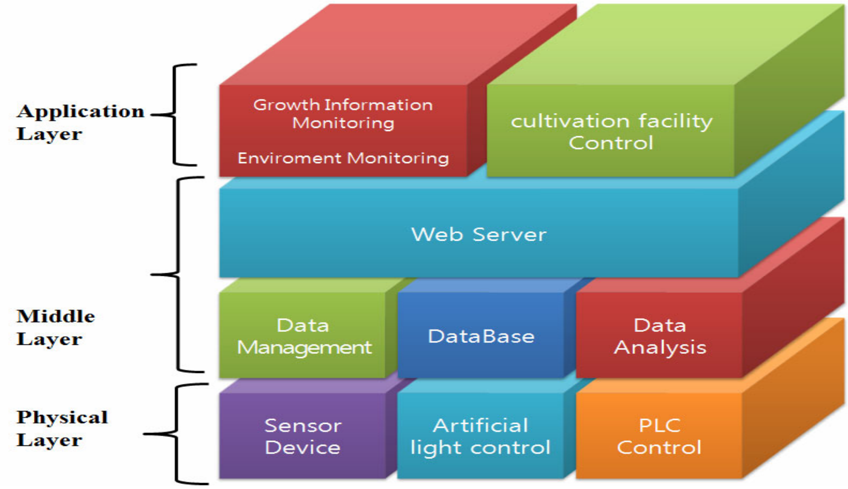


Fig. 3. Paprika Green House System Structure

3.1 System Configuration

‘Paprika green house system’ (PGHS) is comprised of following three layers. The physical layer has ‘environmental sensor’, collecting environmental information, ‘artificial light-source growth control device’ and PLC. The middle layer has data analysis and system control. The application layer has GUI and control.

3.1.1 Physical Layer

Physical layer has a sensor device which collects information from paprika culture media and sends the raw data to middle layer. It also has ‘artificial light-source growth control device’ which controls the wavelength and light quantity of LED light-source for the most optimum growth of paprika. PLC controller controls the temperature, humidity and growth-environment at the root zone based on the collected environmental information. It collects environmental information of cultivation location and makes a control platform to perform the monitoring and control of ‘Paprika green house system’ (PGHS) through each module.

3.1.2 Middle Layer

The middle layer is comprised of ‘data filtering module’, ‘data analysis module’, ‘environment control module’, ‘artificial light-source control module’, ‘database’ and ‘web server’.

‘Data filtering module’ processes the raw data transmitted from sensors and saves the temperature, illuminance, humidity and root zone environment data in the database. ‘Data analysis module’ analyzes the cultivation location environment and crop status based on the information saved in the database. ‘Environment control module’ transmits the control signal to the PLC of physical layer. ‘Artificial light-source control module’ transmits the control signal to the artificial light-source controller. ‘Database’ saves the environment data and analysis data of paprika cultivation location. ‘Web server’ distributes the service to users through WEB-GUI.

3.1.3 Application Layer

The application layer is comprised of WEB-GUI, which provides user with service from ‘Paprika green house system’ (PGHS).

3.2 Service Provided

There are ‘paprika growth information monitoring service’, ‘growth environment monitoring service’, ‘root zone environment monitoring service’ and ‘artificial light-source control service’ and ‘cultivation environment control service’. Details of them are as following.

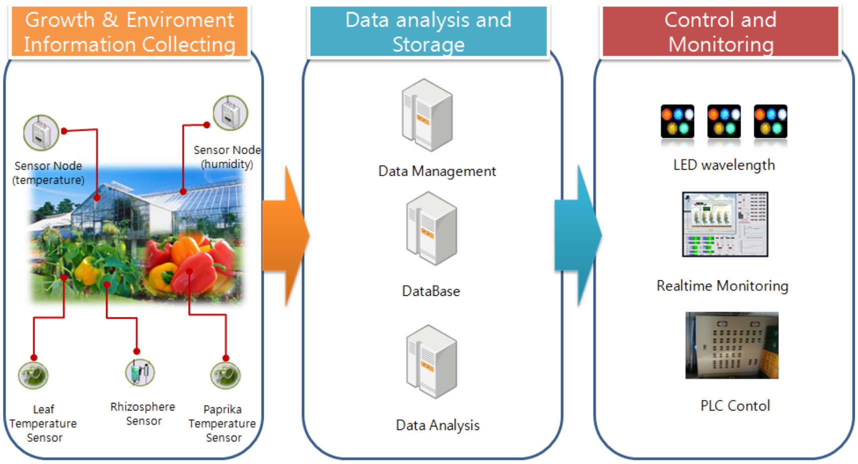


Fig. 4. Providing Service of Paprika Green House

3.2.1 'Growth Environment Monitoring Service'

'Growth environment monitoring service' provides with paprika growth information and Greenhouse information. During paprika cultivation, if the temperature difference between paprika and atmosphere would be more than 4oC, there will be dew condensation and paprika will suffer diseases like grey fungus. To cope with this problem, sensors to measure the temperature of paprika fruit and paprika leaves are located at the fruit surface and within 5cm of the leaves rear-side. The sensors will collect the temperature data in 2 minutes cycle. User can know the temperature difference between paprika and atmosphere through these sensors and can cope with the temperature difference problem caused by temperature difference between crop and atmosphere, in an active way.

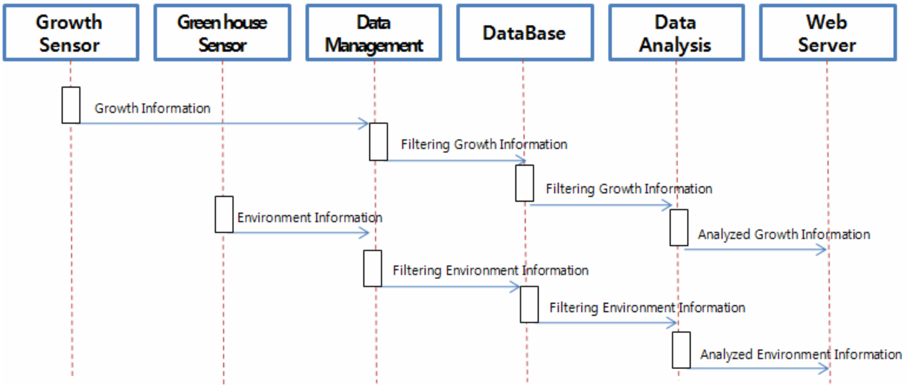

Growth Sensor
Green house Sensor
Data Management
DataBase
Data Analysis
Web Server
Growth Information
Filtering Growth Information
Filtering Growth Information
Environment Information
Analyzed Growth Information
Filtering Environment Information
Filtering Environment Information
Analyzed Environment Information

Fig. 5. Growth Environment Monitoring Service Sequence Diagram

Additional sensors for temperature, humidity and illuminance are installed in the greenhouse so that the user can know the situation through web and identify the growth environment of paprika culture media.

The activation process of ‘growth environment monitoring service’ is as following. Sensors installed in the greenhouse and on the crops collect the raw data. Data management will extract the leaf temperature, leaf wetness and greenhouse environment information (temperature, illuminance, humidity) and save them in the database. The saved data are analyzed and provided to user through web server in the form of web page.

3.2.2 ‘Root Zone Environment Monitoring Service’

Paprika cultivation is mostly done by culture medium; therefore, the root zone management is very important because it gives big impact on the absorption efficiency of culture medium. The root zone of the crop means the soil environment, which changes as the growing roots absorb nutrient and save it. EC and PH are especially important. If EC in soil is not enough, fruit does not grow compared to the leaves growth. If EC is too high, production decreases. If the fertilizer content in the soil becomes higher, PH decreases [6].

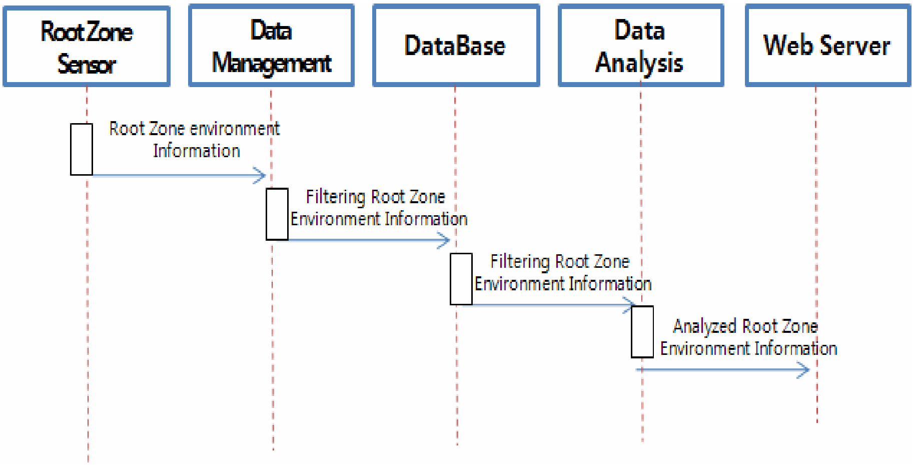

Root Zone Sensor
Data Management
DataBase
Data Analysis
Web Server
Root Zone environment Information
Filtering Root Zone Environment Information
Filtering Root Zone Environment Information
Analyzed Root Zone Environment Information

Fig. 6. Root Zone Environment Monitoring Service Sequence Diagram

This service utilizes such characteristics of EC and PH and monitors the root zone environment. The activation process is same as ‘growth environment monitoring service’.

3.2.3 ‘Artificial Light-Source Control Service’

LED can save 80% energy than existing incandescent light bulbs. LED prevents vermin and adjusts the growth velocity of crops so that shipping timing can be adjusted. As seen in following table, the wavelengths of LED give various impacts on the crops. This service applied those impacts [7] to the ‘artificial light-source control

Table 1. Artificial light-source Impact

Wavelengths	Impact
1400~1000 (IR-A)	No specific impact on crops. Gives heat impact
780(IR-A)	Promotes specific elongation effect on crops
660(red)	Maximize chlorophyll reaction (655)
610(red yellow)	Not good for photosynthesis. Prevents vermin (580~650)
430~440(blue)	Maximize photosynthesis (430), Maximize chlorophyll reaction (440), Entice vermin.
400 ~ 315(UV-A)	In general, makes leaves thick. Encourages the coloring of pigments. Entice vermin.
280(UV-B)	Important reaction in many synthetic processes (makes antibody), Harmful if too strong
100(UV-C)	Let crops wither rapidly.

service' and let it contribute in the control of paprika growth-speed and quality improvement.

3.2.4 'Cultivation Environment Control Service'

'Cultivation environment control service' controls the devices installed in the paprika greenhouse based on the data collected from sensors and saved in the database. The service maintains the optimum environment for the growth of the crop.

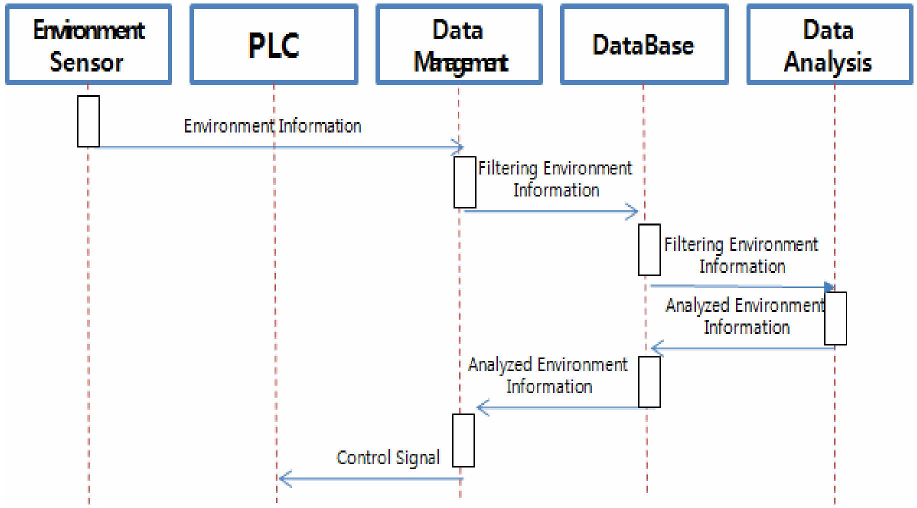

Environment Sensor
PLC
Data Management
DataBase
Data Analysis
Environment Information
Filtering Environment Information
Filtering Environment Information
Analyzed Environment Information
Analyzed Environment Information
Control Signal

Fig. 7. Cultivation Environment Control Service Sequence Diagram

The activation environment is as following. The environmental information sent from cultivation location is transmitted to the data management in the middle layer. They are saved in the database after the correction of overlapping or wrong data. Then the saved data is sent to the ‘data analysis module’, where optimum control information for paprika growth would be analyzed. That information is saved again in the database and signals will be sent to PLC so that it would automatically control the devices such as ventilators and fan heaters.

4 Implementation

Various devices such as sensors, ventilators and fan heaters were installed in the paprika greenhouse to examine the performance capability of ‘Paprika green house system’ (PGHS).

Figure 8 and Figure 9 are sensors measuring the fruit temperature, leaf temperature and leaf wetness.

Figure 10 is sensors collecting the root zone environment information.

Figure 11 is Web-GUI, providing users with ‘Paprika green house system’ (PGHS) service. Information collected from paprika fruits and leaves in Figure 8, Figure 9 can be confirmed in (b) of Figure 11. Root zone information collected in Figure 10 can be confirmed in (c) of Figure 11.

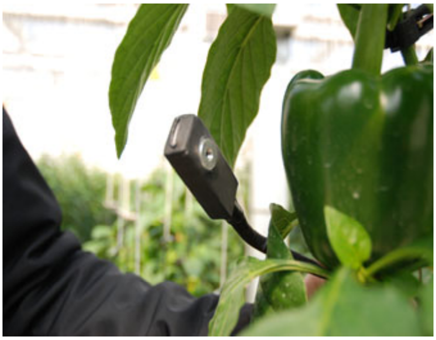


Fig. 8. Fruit Sensor

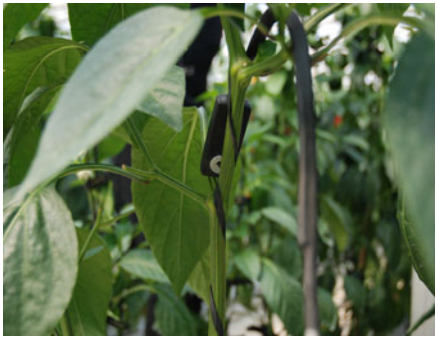


Fig. 9. Leaf Sensor

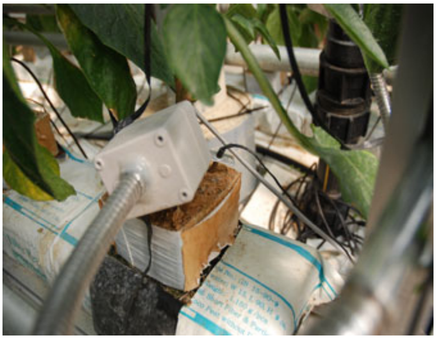


Fig. 10. Root Zone Sensor

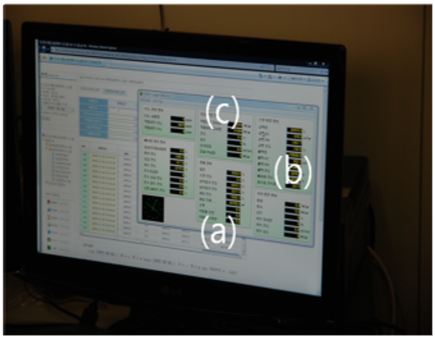


Fig. 11. Web-GUI

The greenhouse environment information values measured in Figure 12 can be confirmed in (a) of Figure 11. Data collected at the sensors go through server and the most optimum growth data will be sent to Figure 13. Then the server sends signals to artificial light-source controller of Figure 13 and PLC controller of Figure 14.

It has been demonstrated that the 'Paprika green house system' (PGHS) shows optimum control performance for the best growth of paprika.

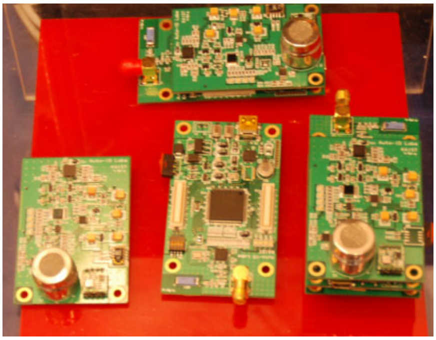


Fig. 12. Environment Sensor

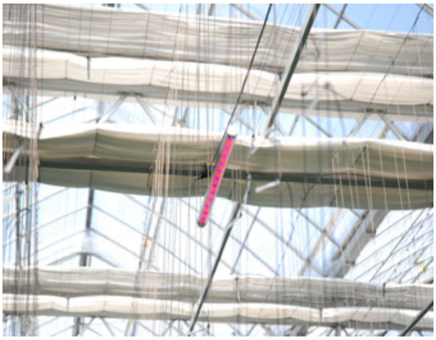


Fig. 13. Artificial Light Controller

Fig. 14. PLC controller

5 Conclusion

This research paper realized the paprika greenhouse environment monitoring system for the precise growth-management of high added-value crop, paprika. The suggested system 'Paprika green house system' (PGHS) makes a network comprised of sensors measuring temperature, humidity, illuminance and others. The system also controls ventilators, humidifiers, lightings and video-processing through web-based GUI by analyzing the measured data.

The suggested 'Paprika green house system' (PGHS) will contribute in the farmers' income increase, which is the top priority task in domestic agricultural industry. The system will create other research results by providing with the growth environment information from numerous paprika cultivation locations. Enhancement of price competitiveness of domestic horticultural industry can also be achieved by improving the distribution rate of new paprika species.

As the next research subject, the occurring conditions of paprika diseases and vermin will be made into database and a ‘diseases and vermin forecast system’ will be designed and implemented, so that farmers can actively cope with the diseases and vermin in advance.

Acknowledgements. This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2010-(C1090-1021-0009)).

References

1. Yooun-il, N.: Present Status and Developmental Strategy if Protected Horticulture Industry in Korea. The KCID Journal 10(2), 191–199
2. Dorais, M.: The use of supplemental lighting for vegetable crop production: light intensity, crop response, nutrition, crop management, cultural practices. In: Canadial Greenhouse Conference (2003)
3. `http://jindo.jares.go.kr/`
4. Lee, E.-J., Lee, K.-l., Kim, H.-S., Kang, B.-S.: Development of Agriculture Environment Monitoring System Using Integrated Sensor Moudle. Korea Contents Association 10, vol. 10(2).
5. Lee, M.-H., Shin, C.-S., Jo, Y.-Y., Yoe, H.: Implementation of Green House Integrated Management System in Ubiquitous Agricultural Environments. Journal of KIISE 27(6), 21–26 (2009)
6. `http://www.kati.net/index.jsp`
7. `http://cafe.naver.com/chled`
8. Jeong, W.-J., Myoung, D.-J., Lee, J.-H.: Comparison of Climatic Conditions of Sweet Pepper’s Greenhouse between Korea and the Netherlands. Journal of Bio-Environment Control 18(3), 244–252 (2009)

A Smart Service Model Using Smart Devices*

Yongyun Cho and Hyun Yoe

Information and Communication Engineering, Sunchon National University,
413 Jungangno, Suncheon, Jeonnam 540-742, Korea
{yycho,yhyun}@sunchon.ac.kr

Abstract. Recently, in smart places or ubiquitous computing environments, there are many researches for smart services using smart devices with sensors, interactive I/O, and convenient UI. This paper propose an smart service model based on smart devices, especially smart phones, in various smart spaces, including urban computing and ubiquitous computing environments. The suggested service model offers an editing UI based on a context-aware workflow model to develop smart services. And, with the suggested model, users can easily uses real data from USN/RFID in various smart spaces as contexts for smart services according to a predesigned ontology. So, in the smart spaces, anyone who is with smart devices can easily make a smart service or application by using the suggested service model.

1 Introduction

The main interests in many of the researches for a smart service may have been how to make the service process automatic without human's interventions [1,2,3,7,8,9]. Recently, the mobility of a smart service, which has to be served regardless of time and place, seems to be considerable. A smart service model in mobile devices and, furthermore, smart devices may be more attractive to match the demand. Because the recent smart devices mostly include smart technologies such as various sensors, more application, and powerful H/W and S/W resources, a research for the smart service model with the smart devices seems to be reasonable and potential in various computing environments. And, because users with smart devices are on the increase more and more, the demands for the smart service using smart devices will be more expanded.

There are many researches for smart services, which have successfully be adopted in such the various smart service domains as u-health, u-home, u-office, and u-city, u-agriculture [1,2,3,4,5,6,7,8,9]. Many of the researches have mainly concentrated in the method how efficiently to represent the real data as contexts and to make the devices in the real world understand the contexts.

* This research was supported by the MKE(Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Advancement)" (IITA-2009-(C1090-0902-0047)).

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 661–666, 2010.

In this paper, we introcude a smart service model for smart devices based in context-aware workflow model in ubiquitous computing environments. The suggested service model offers a convenient editing environment, which is based on a context-aware workflow service model and a context model [1,2,3], to users with smart devices to make a smart service. This paper is constructed as follows: Section 2 describes the related works about context-aware service models or service systems using mobile devices and smart phones. Section 3 describes the conceptual architecture of the suggested smart service model. Section 4 introduces the experiments and the results with an android smart phone to implement the suggested service model. Section 5 mentions the conclusion and the future works.

2 Related Work

Commonly, the studies for smart services with mobile devices or smart devices may be interested in how to support the convenient service without human interceptions and the instant service anywhere and anytime.

CoMeR [7] introduces a N x M-dimensional model and a service system architecture of a hybrid processing approach to use context and profiles to recommend media contents and information for users through a smart phones. CoMeR may be valuable in the view of consideration the various context information, ranging from user preference and situation to devices, as constraint information for the recommendations.

The DYNAMOS [8] is a hybrid research project to use context information for peer-to-peer social functionalities. Especially, the DYNAMOS introduces a system platform and application prototype with smart phones.

Contory [9] is a middleware to be designed for context provisioning on mobile devices. The system introduces the concept, which represents the 3-tier provisioning architecture, which consists of internal sensors-based, external infrastructure-based, and distributed provisioning in ad hoc networks. So, Contory seems to be valuable in the points that it can support a context provisioning, not only flexible and adaptive, but also multiple.

[10] is a recent interesting research about an urban computing management system using mobile phone, which is designed with wireless ad hoc networks. The system includes complete software which can make all kinds of autonomous devices communicate with each others in the urban computing environments. uFlow [11] is a web service-based framework to support a context-aware service using uWDL [2], which is a context-aware workflow language. uFlow can express independent services as a context-aware service flow and provide the functionalities to select an appropriate service based on high-level contexts, profiles, and events information, which are obtained from various sources and structured by ontology [11].

3 A Smart Service Model with Smart Devices

Figure 1 shows the conceptual architecture for the suggested smart service model with smart devices in ubiquitous computing environments.

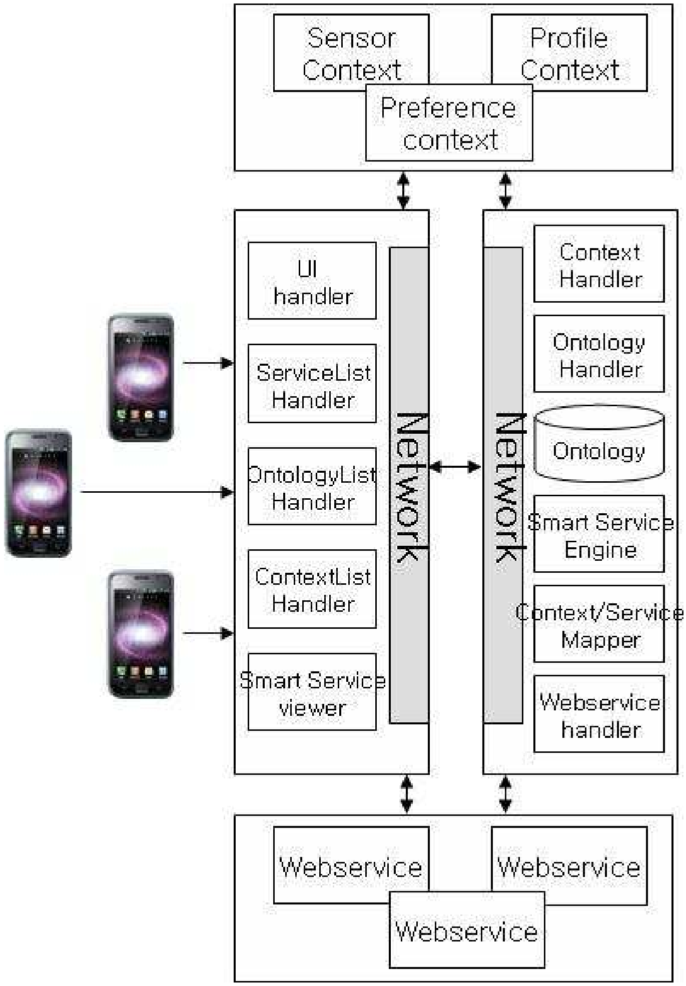


Fig. 1. A conceptual architecture of the suggested smart service model

In Figure 1, the architecture may be divided in two parts, a client side with mobile devices and a service side with a smart service engine. The client side consists of many list handlers to represent the list information of contexts, ontologies, and webservices, which are precessed or stored in the server side, in users' smart devices. The users can conveniently use the offered lists to compose their smart services by themselves. And, the server side has many handlers, which consists of a context handler, an ontology handler, a smart service engine, and a webservice hander. The a context handler is to aggregate low-level contexts from various sensors in real world, and the webservice hander is to manage an open smart service implementation described in a webservice standard service protocol, for example in WSDL. The ontology handler is to manage ontolgies for various service domain. The engine is to process the smart service transferred from the user side according to the contexts.

4 Experiments and Results

In this section, we will do the experiment to compose a smart service with a smart device efficiently and conveniently. To do this, this paper implements a workflow-based smart service executed with an android smart phone. For the implementation, we uses a pentium 4 PC mounted an android 2.0 and eclipse. Figure 2 and Figure 3 show the results of the experimental implementation of the suggested smart service with a smart phone. First, Figure 2 shows the sample contexts listed from pre-defined ontology in an android smart phone.

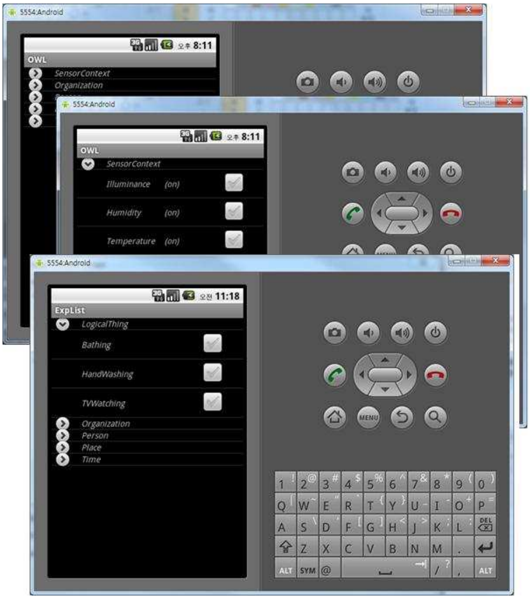


Fig. 2. A sample contexts for possible domains in various smart spaces

In Figure 2, the sample contexts can be categorized and instantly be listed from the ontology defined in OWL [11] according to the domains of the services, which users want to receive. And, the context list list has a hierachical structure, consisting of sub-context items Figure 3 shows a simple RDF-based input window to compose context information with the ontology, and an available service list.

In Figure 3, the window is based on a context description model, which can define a context according to a rule-based context model consisting of the triplet of <subject>, <verb>, and <object> based in RDF [12].

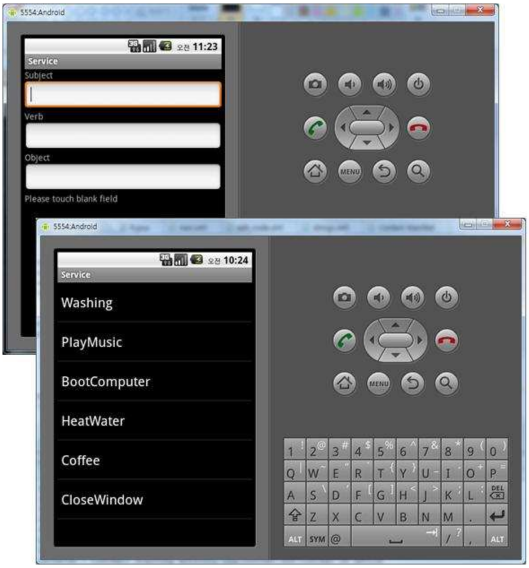


Fig. 3. A simple RDF-based input window and a service list

5 Conclusion

This paper introduced a smart service model with smart devices for various computing environment. The service model supports that users with smart devices can make their smart services, using the open webservice and ontologies through a client/service architecture. For this, the suggested service model has the convenient GUI-based window to make a smart service in client side, which consists of various list handlers. In the experiment, we implemented the suggested client-side, including the various lists transferred from the server side. With the suggested model, users with smart device can conveniently make a smart service, because they can use the domain-directed ontology information when they want a smart service for a specific service domain. In the future work, we will concentrate to the researches to design and implement full GUI-based environment for composing of smart service with smart devices, and to support the type of the plug-in with valuable open APIs for smart service with smart devices.

References

1. Cho, Y., Yoe, H., Kim, H.: CAS4UA: A Context-Aware Service System Based on Workflow Model for Ubiquitous Agriculture. In: Kim, T.-h., Adeli, H. (eds.) AST/UCMA/ISA/ACN. LNCS, vol. 6059, pp. 572–585. Springer, Heidelberg (2010)

2. Han, J., Cho, Y., Choi, J.: Context-Aware Workflow Language based on Web Services for Ubiquitous Computing. In: Gervasi, O., Gavrilova, M.L., Kumar, V., Laganá, A., Lee, H.P., Mun, Y., Taniar, D., Tan, C.J.K. (eds.) ICCSA 2005. LNCS, vol. 3481, pp. 1008–1017. Springer, Heidelberg (2005)
3. Cho, Y., Moon, J., Yoe, H.: A Context-Aware Service Model Based on Workflows for u-Agriculture. In: Gervasi, O. (ed.) ICCSA 2010, Part III. LNCS, vol. 6018, pp. 258–268. Springer, Heidelberg (2010)
4. Tang, F., Guo, M., Dong, M., Li, M., Guan, H.: Towards Context-Aware Workflow Management for Ubiquitous Computing. In: Proceedings of ICESS 2008, pp. 221–228 (2008)
5. Dey, A.: Understanding and Using Context. Personal and Ubiquitous Computing 5(1) (2001)
6. Choi, J., Cho, Y., Choi, J.: The Design of a Context-Aware Workflow Language for Supporting Multiple Workflows. Journal of Korean Society for Internet Information 11(1), 145–158 (2009)
7. Yu, Z., Zhou, X., Zhang, D., Chin, C.-Y., Wang, X., Men, J.: Supporting Context-Aware Media Recommendations for Smart Phones. IEEE Pervasive Computing 5(3), 68–75 (2006)
8. Riva, O., Toivonen, S.: The DYNAMOS approach to support context-aware service provisioning in mobile environments. Journal of Systems and Software 80(12), 1956–1972 (2007)
9. Riva, O.: Contory: a middleware for the provisioning of context information on smart phones. In: van Steen, M., Henning, M. (eds.) Middleware 2006. LNCS, vol. 4290, pp. 219–239. Springer, Heidelberg (2006)
10. Mitra, K., Bhattacharyya, D., Kim, T.: Urban Computing and Informaiton Management System Using Mobile Phone in Wireless Sensor Network. Internation Journal of Control and Automation 3(1), 18–26 (2010)
11. Lauser, B., Sini, M., Liang, A., Keizer, J., Katz, S.: From AGROVOC to the Agricultural Ontology Service / Concept Server - An OWL model for creating ontologies in the agricultural domain. In: Proceedings of the OWLED 2006 Workshop on OWL: Experiences and Directions, Athens, Georgia, USA, pp. 10–11 (2006)
12. Han, J., Cho, Y., Kim, E., Choi, J.: A Ubiquitous Workflow Service Framework. In: Proceedings of the 2006 International Conference on Computational Science and its Application, pp. 30–39 (2006)

An Implementation of the Salt-Farm Monitoring System Using Wireless Sensor Network

JongGil Ju, InGon Park, YongWoong Lee, JongSik Cho, HyunWook Cho, Hyun Yoe, and ChangSun Shin*

Dept. of Information and Communication Engineering,
Sunchon National University, Korea
{jake,crescent1,cho1318,chohyunwook,pig9004,
hyoe,csshin}@sunchon.ac.kr

Abstract. In producing solar salt, natural environmental factors such as temperature, humidity, solar radiation, wind direction, wind speed and rain are essential elements which influence on the productivity and quality of salt. If we can manage the above mentioned environmental elements efficiently, we could achieve improved results in production of salt with good quality. To monitor and manage the natural environments, this paper suggests the Salt-Farm Monitoring System (SFMS) which is operated with renewable energy power. The system collects environmental factors directly from the environmental measure sensors and the sensor nodes. To implement a stand-alone system, we applied solar cell and wind generator to operate this system. Finally, we showed that the SFMS could monitor the salt-farm environments by using wireless sensor nodes and operate correctly without external power supply.

Keywords: Solar Salt, USN, Salt-farm, Environment Monitoring System, Renewable Energy.

1 Introduction

The information technologies, wireless sensors, ubiquitous computing and communication devices techniques are applied to various industrial fields. But solar salt industry didn't receive above IT technologies yet. The solar salt is very sensitive to salt-farm environments. If we collect precision salt-farm environment data, the productivity and quality rate of products will be improved.

For producing high quality salt, we propose the Salt-Farm Monitoring System (SFMS) that uses hardware and software IT technologies. This system monitors and collects the information of salt-farm environments with a renewable power supply.

The rest of this paper is organized as follows: Section 2 is related works, Section 3 describes the system architecture of the SFMS and Section 4 presents implementation of the system. Finally, we discuss conclusions and future works in Section 5.

* Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 667–675, 2010.

2 Related Works

There are five requisites for growing corps [1]. They are temperature, light, air, water and soil. A project of plant growth monitoring system was developed by Go-heung agriculture technology & extension center in Korea. They applied environment monitoring sensors and software to Hanabong farm [2]. This system measures the greenhouse environment status by using sensors adhered to plant and sends information to grower's home via internet.

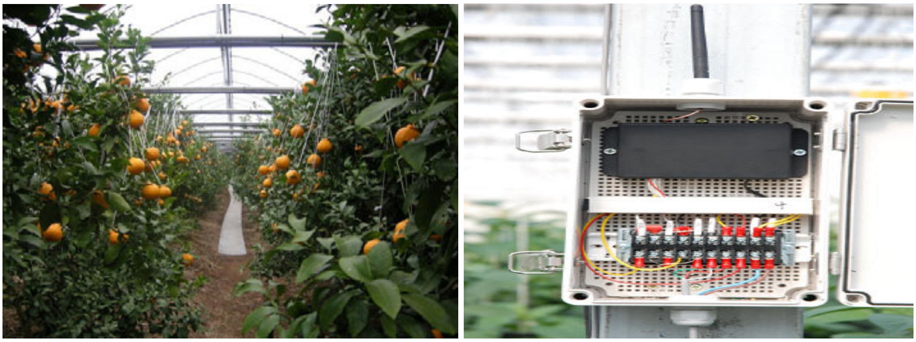


Fig. 1. The Hanabong monitoring system by Go-heung agriculture technology & extension center

Floating buoy is developed to monitor ocean environments via Orbcomm satellite and a method is proposed to increase measurement accuracy of sea water temperature with common low price temperature sensor [3].

The feasibility of the developed node was tested by deploying a simple sensor network into Martens Greenhouse Research Foundation's greenhouse in Närpiö town in Western Finland. They are the number of wireless sensor networks with tree structure form integrating sensors of the same category. In addition, three commercial sensors capable to measure four climate variables [4, 5].

By using above technologies, agricultural sensors and nodes are proposed, and various applications are developed.

3 Salt-Farm Monitoring System (SFMS) Architecture

The Salt-Farm Monitoring System (SFMS) is a collect real time environmental field data from various sensors. Figure 2 shows the architecture of the SFMS.

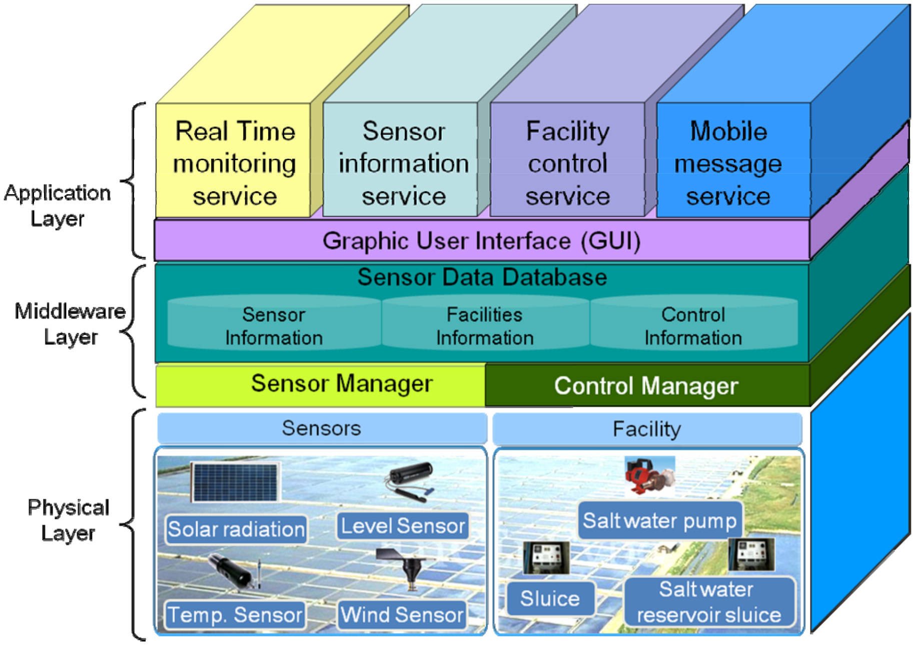


Fig. 2. Salt-Farm Monitoring System's architecture consisting of three layers

The SFMS divided into three layers. The physical device layer includes sensors and facilities. Sensors are temperature, humidity, solar radiation, wind direction, wind speed, rain and salt water level sensor. Facilities are salt water reservoir sluice, salt-farm sluice and salt water reservoir pump. SFMS is a self-charging stand alone system using renewable energy. We supply solar-power, wind generator into the system without any electric power.

The middle layer has the sensor manager, control manager, salt-farm database. The sensor manager manages the information from environments sensors [6]. The control manager controls the facilities device using the salt-farm database. The sensor data database provides environment information from physical layer devices to application layer via sensor, facilities control information.

The application layer provides with the real time monitoring service, sensor information service, facility control service and mobile message service [7]. These are provided with laptop GUI (Graphic User Interface), web GUI and mobile phone GUI. Three layers are integrated into the SFMS. By interacting with each layer, the system provides users with salt-farm environment information.

3.1 Environment Monitoring Service

The salt-farm should be real-time sensing, because it is very sensitive to environment element such as temperature, humidity, illumination, etc.

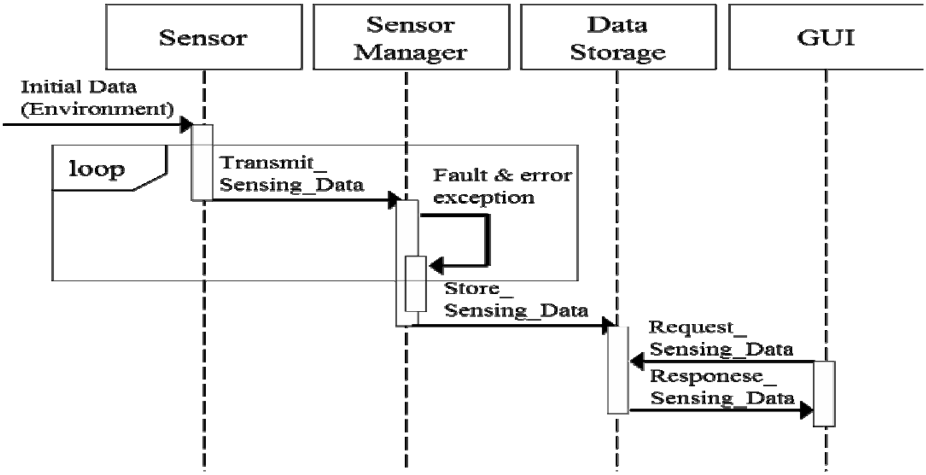


Fig. 3. Environmental Information Service sequence diagram

Figure 3 shows procedure of the environment monitoring service. First, this service sends the raw data of environment sensors to the sensor manager. The raw data are temperature, humidity, salinity and intensity of illumination information. Sensor manager verify which the raw data is error data or not. Wireless sensor network has virtually the low reliability in an open area. In order to reduce the risk of collecting environment information it needs to data filtering. A simple filtering such as compared with previous raw data through the classification and then stored in the database. A GUI obtained the sensor data from data storage then offers the environment information to users.

3.2 Control Services

Figure 4 shows procedure of automatic control services facilities. Database provides the factors information such as facilities state and sensing data to GUI. GUI send the

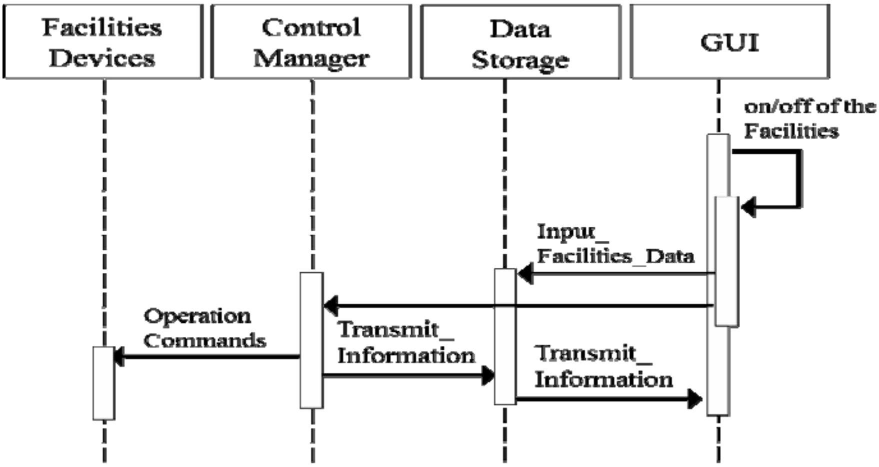


Fig. 4. Facilities Automatic Control Services Sequence Diagram

on/off command signal to control manager through database by Logical analysis of response factors data [8].

4 Implementation of the SFMS

In this chapter, we implement the SFMS by implementing the system's components. Figure 5 is the system model.

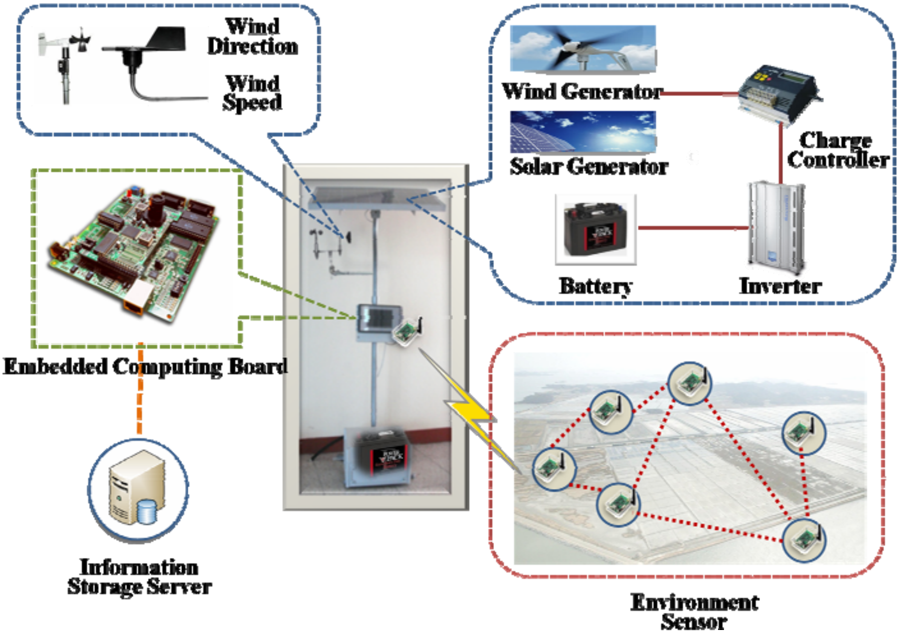


Fig. 5. SFMS model including embedded board, renewable charging devices and sensors

The SFMS was applied a renewable energy system. The system has solar cell, wind generator and storage battery. The system stores power in the daytime and using in the night time.

4.1 System Components

This system includes of physical devices and software modules. The physical devices have sensing and information gathering devices. You can see the devices in Figure 6.

Table 1 is showing the power consumption of equipped modules and power supply of solar and wind in the SFMS. The total power consumption of equipped modules like embedded board, salinity sensor and environment sensor is 11.95W. The solar cell and wind generator supplies with electric power of the maximum 200W for each in the 25℃ test environment. This is enough to operate the SFMS. You can see the main system installed sensors' data receiver and database in Figure 7 and a charging battery in Figure 8.

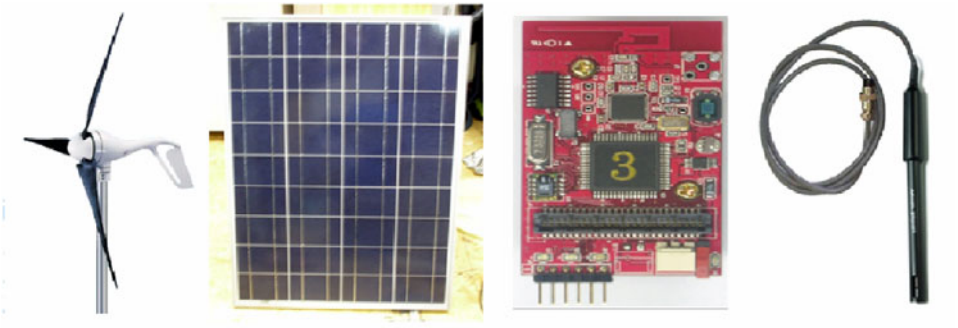


Fig. 6. Wind turbines, Solar cell, Network sensor node, temperature sensor in the SFMS

Table 1. Power consumption of each module and Power supply of charging battery

Module	Power consumption		
	Voltage	Current	Power
Embedded Board	DC 9V	500mA	5W
Environment Sensor	DC 3V	2.3A	6.9W
Salinity Sensor	DC 5V	10mA	0.05W
TOTAL	DC 17V	2.81A	11.95 W
Module	**Supply Power**		
	Voltage	Current	Power
Solar Cell	DC 26.4V	7.6A	200W
Wind generator	DC 24V	7.7A	200W
	Avr Wind Speed 12.5m/s		
Battery	Voltage	Capacity(20HR)	
	DC 12V	64A	

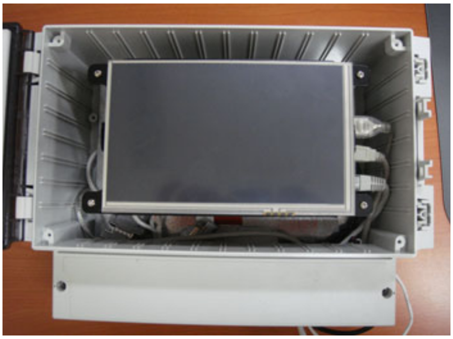


Fig. 7. Embedded board including environment sensor receiver and database

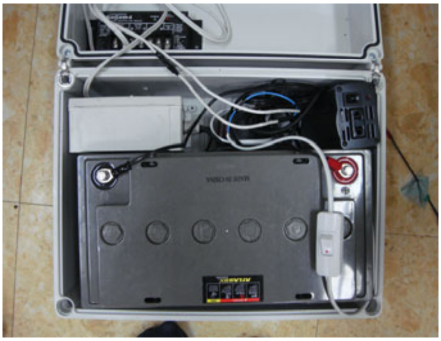


Fig. 8. salinity sensor receiver, integrated battery

Now, we integrate above components into the system. Figure 9 shows the SFMS's prototype including the software modules. The SFMS can apply various environments such as precision agriculture, aquaculture, fishing industry, livestock industry, greenhouse monitoring and salt farm monitoring.

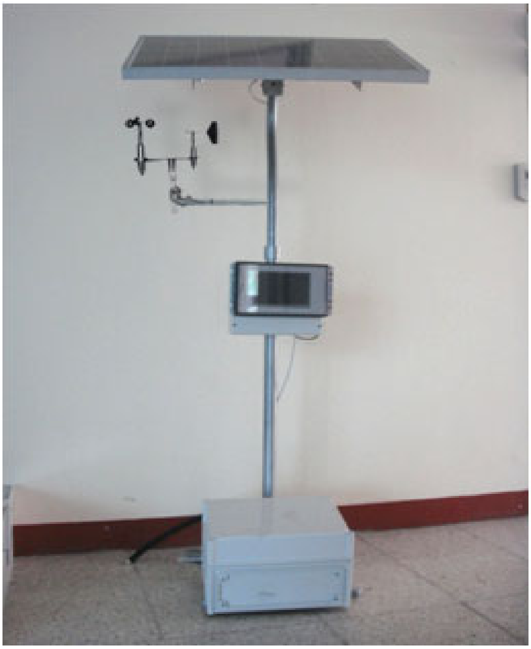


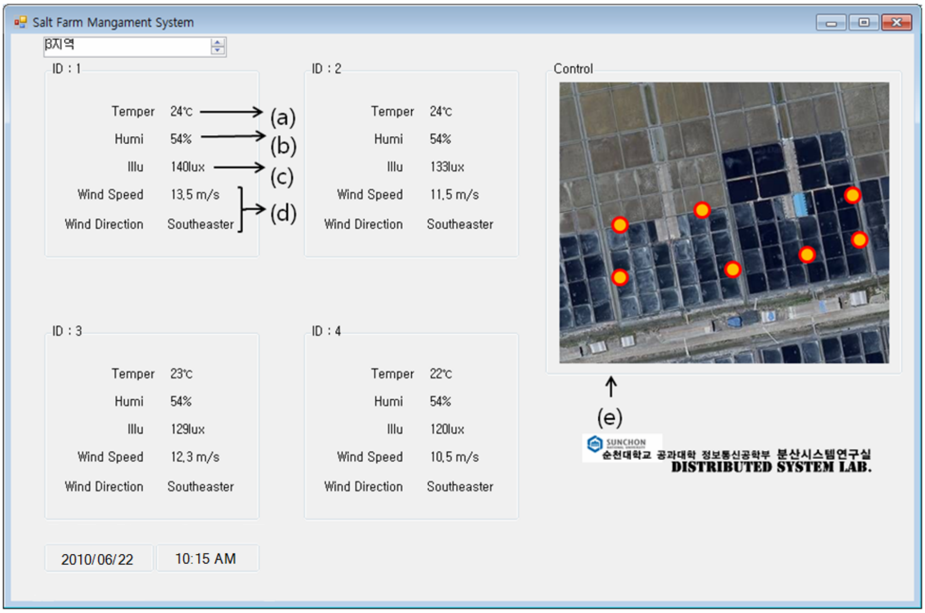
Fig. 9. Prototype of the SFMS

4.2 Implementation Results

Figure 10 is the SFMS's GUI. The (a) shows the sensing value from the temperature sensor. The (b) is sensing value for humidity. The (c) showing the solar radiation sensing value, and the (d) is the sensing value from the wind speed, direction sensors. The (e) shows control of floodgate.

To confirm the successful operation of the SFMS using self-supply of electric power, we perform field test on a sunny day with a mean temperature of 25□ degree and wind speed 12.5m/s. As a result, during the daytime solar cell, wind generator generated power together and night time only wind generator generated the power. If there is windless and cloudy day, SFMS could be supplied power by recharged battery.

Hence, our SFMS can operate with the support of solar cell and wind power in the field without power supply from wired link or additional recharging process. Figure 11 shows a graph of field test result in power consumption.



Salt Farm Mangament System

B지역

ID : 1

Temper 24℃ (a)

Humi 54% (b)

Illu 140lux (c)

Wind Speed 13.5 m/s (d)

Wind Direction Southeaster

ID : 2

Temper 24℃

Humi 54%

Illu 133lux

Wind Speed 11.5 m/s

Wind Direction Southeaster

Control

(e)

ID : 3

Temper 23℃

Humi 54%

Illu 129lux

Wind Speed 12.3 m/s

Wind Direction Southeaster

ID : 4

Temper 22℃

Humi 54%

Illu 120lux

Wind Speed 10.5 m/s

Wind Direction Southeaster

SUNCHON 순천대학교 공과대학 정보통신공학부 분산시스템연구실 DISTRIBUTED SYSTEM LAB.

2010/06/22 10:15 AM

Fig. 10. FMSS's GUI

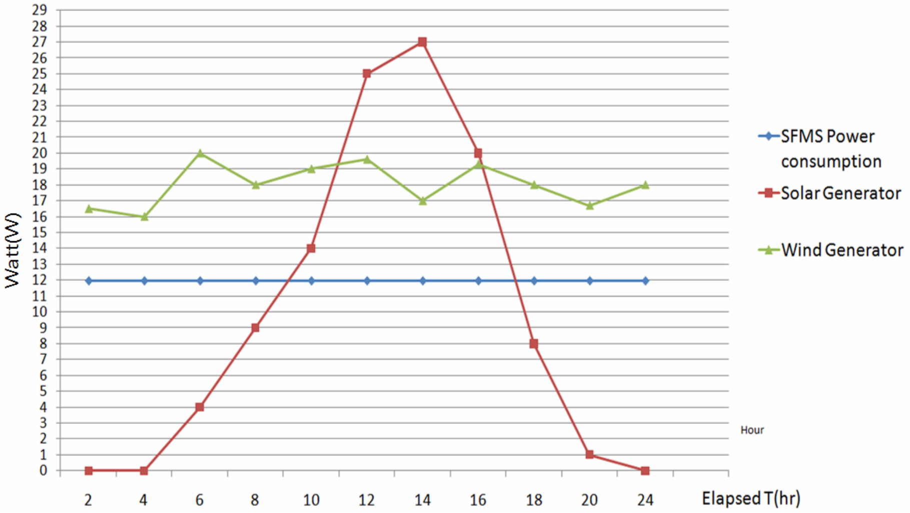


Fig. 11. Field test result of generate power and consume power

5 Conclusions

This paper proposed the Salt-Farm monitoring service (SFMS) that could monitor environments of salt-farm using renewable energy. Also, for verifying the execution of our system, we implemented system's components and made the SFMS prototype. Then we showed the executing results of the system. From this result, we confirmed that our system could monitor the salt-farm conditions by using various sensors and facilities. Also, we show that renewable energy can make operating the SFMS without any external power.

For future works, we aim to developing an improved monitoring system which operates based USN and applies into the salt storage inventory. Also, it's challenge for us to keep a good condition from salt wind, salt water and extreme weather.

Acknowledgements. This research was supported by the MKE (Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Advancement) (IITA-2009-(C1090-0902-0047)).

References

1. Shin, C.S., Joo, S.C., Lee, Y.W., Sim, C.B., Yoe, H.: An Implementation of Ubiquitous Field Server System Using Solar Energy Based on Wireless Sensor Networks. Studies in Computational Intelligence 209 (2009)
2. Lee, Y.W., Cho, C.J., Ju, J.K., Shin, C.S., Lee, J.H., Shin, H.H., Yum, Y.C., Yoe, H.: Implementation of System for a Ubiquitous Farming-diary. Journal of the Korean Society of Agricultural Engineers 52(2), 35–42 (2010)
3. Yu, Y., Gang, Y., Lee, W.: Development of a Floating Buoy for Monitoring Ocean Environments. Journal of the Korean Society of Marine Engineering 33 (2009)
4. Sensinode. OEM Product catalog (2007), http://www.sensinode.com/pdfs/sensinode-catalog-20071101.pdf
5. Ahonen, T., Virrankoski, R.: Greenhouse Monitoring with Wireless Sensor Network. In: IEEE/ASME International Conference on (2008)
6. Delin, K.A., Jackson, S.P., Burleigh, S.C., Johnson, D.W., Woodrow, R.R., Britton, J.T.: The JPL Sensor Webs Project: Fielded Technology. In: Space Mission Challenges for IT Proceedings, Annual Conference Series, pp. 337–341 (2003)
7. Shin, C.S., Kang, M.S., Jeong, C.W., Joo, S.C.: TMO-based Object Group Framework for Supporting Distributed Object Management and Real-Time Services. In: Zhou, X., Xu, M., Jähnichen, S., Cao, J. (eds.) APPT 2003. LNCS, vol. 2834, pp. 525–535. Springer, Heidelberg (2003)
8. Kang, B.J., Park, D.H., Cho, K.R., Shin, C.S., Cho, S.E., Park, J.W.: A Study on the Greenhouse Auto Control System based on Wireless Sensor Network. International Conference on Security Technology, pp. 41–44 (December 2008)
9. Tilman, D., Cassman, K.G., Matson, P.A., Naylor, R., Polasky, S.: Agricultural sustainability and intensive production practices. Nature 418, 671–677 (2002)
10. Burrell, J., Brooke, T., Beckwith, R.: Vineyard computing: sensor networks in agricultural production. IEEE Pervasive Computing 3(1), 38–45 (2004)
11. Yoe, H., Eom, K.-B.: Design of Energy Efficient Routing Method for Ubiquitous Green Houses. In: 1st International Conference on Hybrid Information Technology (November 2006)
12. Lee, M.H., Be, K., Kang, H.J., Shin, C.S., Yoe, H.: Design and Implementation of Wireless Sensor Network for Ubiquitous Glass Houses. In: 7th IEEE/ACIS International Conference on Computer and Information Science, pp. 397–400 (May 2008)

Mobile Business Agents Model and Architecture

Haeng Kon Kim

Department of Computer Engineering, Catholic University of Daegu,
Kyungsan, Kyungbuk, 712-702, Seoul of Korea
hangkon@cu.ac.kr

Abstract. Agent-component technology and agent-oriented software engineering have the potential to be more powerful than traditional. Most agent and e-service systems offer several capacities that work together to provide unprecedented flexibility and promise to be more effective at handling the resulting software's evolution and distribution. Therefore, in order to support agent service or agent based business application and system there is the necessity of research about agent development based component. In this paper, we identify and classify the general and e-business oriented agent affecting CBD. We suggest the e-business agent oriented component reference architecture. We also propose systemical development process using AUML(Agent Unified Modeling Language) and design pattern technology to analysis, design and develop e-business agent. Finally we describe how these concepts may assist in increasing the efficiency and reusability in business application and e-business agent development.

Keywords: E-Business Agent, Agent Classification, Component Architecture, CBD, Agent Design Patten.

1 Introduction

Agent-oriented techniques represent an exciting new means of analyzing, designing and building complex software systems. They have the potential to significantly improve current practice in software engineering and to extend the range of applications that can feasibly be tackled. As the demand for more flexible, extensible, and robust Web-based enterprise application systems accelerates, adopting new software engineering methodologies and development strategies becomes critical. These strategies must support the construction of enterprise software systems that assemble highly flexible software components written at different times by various developers[1, 2].

In this paper, we identify the primary and general attribute from existing application and classify the agents form e-business domain as a sub research to develop e-business agent based component. Through all over this, common area is extracted both general agent and e-business agent and e-business agent oriented component with reference architecture. We also propose systemical development process using AUML and design pattern technology to analysis, design, and develop e-business agent. Component reference architecture through agent domain classification is based on component development life cycle. Moreover, the development of e-business agent and system can be obtained the efficiency through component technology.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 676–685, 2010.

2 Related Works

2.1 Basic Characteristics of E-Business Agents

An agent must have a model of its own domain of expertise and a model of the other agents that can provide relevant information. The awareness model of an information agent does not need to contain a complete description of the other agents' capabilities, but rather only those portions that may be directly relevant when handling a request that cannot be serviced locally. In general, we require that intelligent business agents possess distinguishing characteristics described in the following paragraphs[3].

- Delegation abilities: The central idea underlying agents is that of delegation. The owner or user of an agent delegates a task to the agent and the agent autonomously performs the task of behalf of the user. Alternatively, a business agent may decompose the task and delegate parts of it to other agents, which perform the subtasks and report back to the business agent. The agent must be able to communicate with the user or other agents to receive its instructions and to provide results of its activities.
- Agent communication languages and protocols: A business agent is an autonomous entity, hence it must negotiate with other agents to gain access to other sources and capabilities. To enable the expressive communication and negotiation required and organize communications between agents, a language that contains brokering performatives can be particularly useful. Some general examples of agent development environments include the agent builder and the agent library.
- Self-representation abilities: One of the most challenging problems is for agents to express naturally and directly business and system aspects and then combine these into a final meaningful application or implementation. This results in self-describing, dynamic, and reconfigurable agents that facilitate the composition of large-scale distributed applications, by drawing upon business processes and the functionality of existing information sources. Such ideas can benefit tremendously from techniques found in reflection and meta-object protocols.

2.2 AUML(Agent Unified Modeling Language)

The current UML is sometimes insufficient for modeling agents and agent-based systems. However, no formalism yet exists to sufficiently specify agent-based system development. To employ agent-based programming, a specification technique must support the whole software engineering process[4]. Both FIPA(Foundation for Intelligent Physical Agents) and the OMG Agent Work Group are exploring and recommending extensions to UML[1]. The AUML present a subset of an agent based extension to the standard UML for the specification of AIP(Agent Interaction Protocols) and other commonly used agent based notions. An AIP describes a communication pattern as an allowed sequence of messages between agents and the constraints on the content of those messages.

Interaction protocols were chose because they are complex enough to illustrate the nontrivial use of AUML and are used commonly enough to make this subset of AUML useful to other researchers. Agent interaction protocols are a good example of software patterns that are ideas found useful in one practical context and probably useful in others. A specification of an AIP provides an example or analogy that we might use to solve problems in system analysis and design. AUML suggest a specification technique for AIPs with both formal and intuitive semantics and a user-friendly graphical notation. The semantics allows a precise definition that is also usable in the software-engineering process. The graphical notation provides a common language for communicating AIPs[5].

3 E-Business Agent Oriented CBD Reference Architecture

Component reference architecture classifies general agent which analysis existing agent system based on primary property. We can classify the agents by domain attributes, usages and requirement in e-business system. The common area identified from general agent type and e-business agent domain. The reference architecture is constructed based on identified common area and support components for e-business agent development. The architecture offers guideline for adaptable component analysis and design. It also supports component deployment and management.

3.1 Classification

The agents found in systems have special requirements: they must execute as software, hardware, robotics, or a combination of these. Agent developers have identified several forms of agents that are important for application development. The list of agent characteristics presented earlier addresses some of these requirements. Additionally, since agent system has special needs, software and hardware-related forms must be considered. We attempts to palace existing agents into different agent classes. Then, its goal is to construct component reference architecture for e-business agent. We consider both primary attributes and business attributes in existing agent system.

3.1.1 Classification of General Agent

The type of agent, definition, and name, which are used in existing agent system, are circulate in various way. We identify fourteen different types of agents with attribute as in figure 1. We would overview them in terms of some or all of the following.

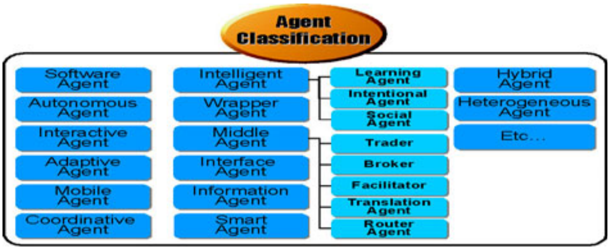


Fig. 1. General Agent Classification

Software agent is defined as *an autonomous software entity that can interact with its environment*. This means that they are autonomous and can react with other entities, including humans, machines, and other software agents in various environments and across various platforms. When an agent has a certain independence from external control, it is considered autonomous. Without any autonomy, an agent would no longer be a dynamic entity, but rather a passive object such as a part in a bin or a record in a relational table. Interactive agents can communicate with both the environment and other entities and can be expressed in degrees. An agent is considered *adaptive* if it is capable of responding to other agents and/or its environment to some degree. At a minimum, this means that an agent must be able to *react* to a simple stimulus, predetermined response to a particular event or environmental signal.

While stationary agents exist as a single process on one host computer, mobile agents can pick up and move their code to a new host where they can resume executing. The rationale for mobility is the improved performance that can sometimes be achieved by moving the agent closer to the services available on the new host. Human organizations exist primarily to coordinate the actions of many individuals for some purpose. Using human organizations as an analogy, systems involving many agents could benefit from the same pattern. Some of the common coordinative agent applications involve supply chains, scheduling, vehicle planning, problem solving, contract negotiation, and product design.

After decades, the term *intelligent* has still not been defined for artificial system and applying the term now to agents may not be appropriate. Most tend to regard the term *agent* and *intelligent agent* as equivalent. Perhaps this is just an attempt to communicate that agents have more power that conventional approaches. Some kinds of intelligent agents are learning agent, intentional agent and social agent. Client agents can relay commands to the wrapper agent and have them invoked on the underlying services. The role provided by the wrapper agent provides a single generic way for agents to interact with non-agent software systems.

Broadly, agentized middleware including agentized common and basic object services. Some kinds of middle agents are trader, broker, facilitator, translation agent and router agent. An interface agent is a program that is able to operate within a user interface and actively assist the user in operating the interface and manipulating the underlying system. An interface agent is able to intercept the input from the user, examine it, and take appropriate action. Agents in the interface can function as a bridge between domain knowledge about the data management systems and the user.

Smart agents are supposed to be able to learn as they react and/or interact with their external environment, so that, with time, their performance increases. Hybrid agents refer to those whose constitution is a combination of two or more agent *philosophies* within a singular agent. The key *hypothesis* for having hybrid agents or architectures is the belief that, for some application, the benefits accrued from having the combination of philosophies within a singular agent is greater than the gains obtained from the same agent based entirely on a singular philosophy. Heterogeneous agent systems, unlike hybrid systems described in the preceding section refers to an integrated set-up of at least two or more agents, which belong to two or more different agent classes.

3.1.2 Agent Classification of e-Business Agent

The current kinds of applications that employ agents is still limited. Once the concepts become more accepted and more tools become available, the agent-based approach will become more embedded it e-business domain and applications. Figure 2 shows an agent classification with e-business attribute and function.

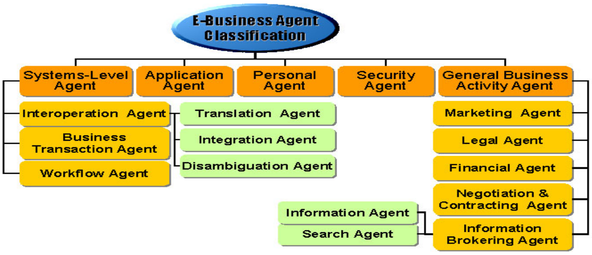


Fig. 2. E-Business Agent Classification

In a e-business environment it is necessary to organize agents into different categories depending on their functionality and competencies. The five basic type of agents can be distinguished as described here. System-level agents exist on top of the distributed objects infrastructure, typically implemented in CORBA by means of the IIOP, which provides objects with transparent access not only to other application objects but also to such facilities as transaction processing, permanent object storage, event services, and the like. Agent solutions are deployed as an extension of the distributed object foundation and may assist in accomplishing the following systems related tasks. Some of the advanced functionality agents required providing support for e-commerce and interoperation of open market business processes are described here. It includes *interoperation agent, business transaction agent, work flow agent.*

A business-to-business e-commerce application is a networked system that comprises a large number of application agents. Each agent is specialized to single area of expertise and provides access to the available information and knowledge sources in that domain and works cooperatively with other agents to solve a complex problem in that vertical domain. This results in the formation of clusters of information sources around domains of expertise handled by their respective agents.

Personal agents work directly with users to help support the presentation, organization, and management of user profile, requests, and information collections. A personal agent gives its user easy and effective access to profile related specialized services and information widely distributed on the Web. The user's agent observes and monitors the actions taken by the user in the interface and suggests better ways to perform the task. These agents can assist users in forming queries, finding the location of data, and explaining the semantics of the data, among other tasks.

The activities and functions of e-business need certain basic agent technology support that is likely to become the basis for developing standard digital agents for e-business. General business agents perform a large number of general commerce support activities that can be customized to address the needs of a particular business organization. It includes *marketing, legal, negotiation, information brokering agent.*

E-business communication need to be guarded by specially designed agents that provide the security services required for the conduct of e-business. Agent support for secure e-business can be segmented into five distinct categories: authentication, authorization, data integrity, confidentiality, and non-repudiation.

- Authentication agents can be used to identify the source of a message sent over the Internet.

- Authorization agents may control access to sensitive information once identity has been verified. Thus, certain transactions may need to be partly accessible to certain parties, while the remainder of the transaction is not. The transaction workflow and authorization agents can coordinate these tasks.
- Secure transactions should guarantee that a message has not been modified while in transit. This is commonly known as integrity and is often accomplished through digitally signed digest codes. Transactions should also guarantee confidentiality.
- Confidentiality refers to the use of encryption for scrambling the information sent over the Internet and stored on servers so that eavesdroppers and interlopers cannot access the data.
- Non-repudiation is of critical importance for carrying out transactions over the Internet. It consists of cryptographic receipts that are created so that the author of a message cannot falsely deny sending a message.

3.2 Component Reference Architecture for E-Business Agent Development

In order to construct component reference architecture, agent is classified in general agent type and e-business function attribute. Figure 3 is a component and meta architecture of based on all above described for e-business agent.

Reference architecture is consisted of dimension, which has 15 general types and 11 concrete business agent types with domain oriented component architecture. These two classification areas tend to be independent for each cross-referenced. Each area has its own horizontal and vertical characteristics. General agent types are corresponding to agent platform and application. It is possible to develop agent system or application by the referencing architecture. The technology of agent can be applied to business domain. Developed component is classified by the reference architecture and is placed according to general agent type and business attribute. In case agent is applied to the agent system or business domain, system is possibly to build up by identifying component related to business domain and combining it.

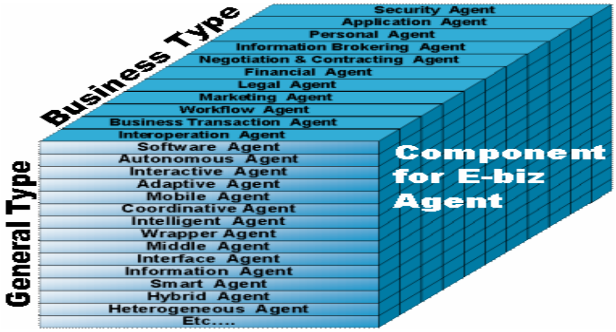


Fig. 3. CBD Reference Architecture of E-Business Agent

4 Agent Component Development Process Based Architecture

As we suggested CBD reference architecture in previous chapter, component development process based architecture is a set of activities and associated results, which lead to the production of a component as in figure 4. These may involve the development of component from UML model.

In figure 4, architecture is at the center of analysis, design, component development, this process applies and designs architecture from early domain analysis phase to component implementation. In addition, we consider systemical development process using AUML and design pattern technology to analysis, design, and develop e-business agent. The domain analysis specification, design model, implemented component, which are produce in process, are stored in the repository[6].

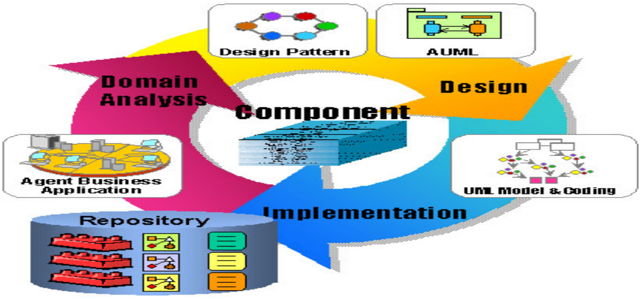


Fig. 4. Component development process

4.1 Agent Domain Analysis Phase

The requirement of agent should be first identified in desired business system. The primary property of agent should be analyzed after that the description for specific agent platform and the sorts of essential properties should be understood. At the same time, it is very important to consider weather the requirement, which is already defined, is corresponding to agent type in reference architecture and what business concept is focused on.

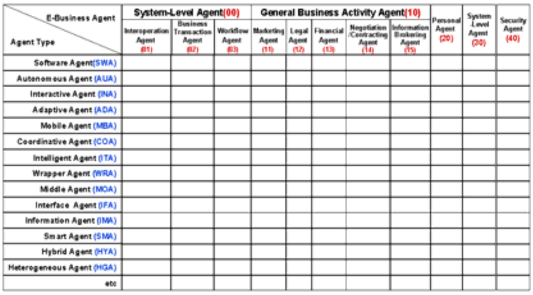
All over those things make high understanding for domain requirement and become referenced to define agent attribute. Selecting of component domain can easily identify design pattern in design phase and easily deploy component. Domain analysis is presented on entire domain concept and scenario using activity diagram. Requirement analysis is defined through use case diagram, and use case description.

4.2 Agent Design Phase

The e-business agent with adaptable component is designed based on domain requirement. Attribute and behavior are defined using class diagram for component, which is expected to be implemented depending on agent type. The definition of component interface is presented on sequence diagram. Contract specification to describe pre-condition, post-condition, and interface properties show the relationship between components. There are two considerations depending on agent property and design technology on design phase. First, part of related to agent interact protocol use AUML notation. And agent interact protocol is described communication pattern. This proposes three levels for the protocols presentation method of agent.

- • Overall protocol level: There are two techniques that best express protocol solutions for reuse; package diagram and templates.
- • Interactions among agents level: There are presented through UML's dynamic model; sequence, collaboration, activity and state diagram.
- • Internal agent processing level: At the lowest level, requires spelling out the detailed processing that takes place within an agent in order to implement the protocol. This layer preset to use activity diagram and state charts.

Second, design pattern can be applied to previously identified area in reference architecture. Figure 5 is design pattern matrix based on meta architecture of component reference architecture and design pattern is identified in matrix. Design pattern is made considering agent functionality and added on other information for component development. Moreover, the concurrency of architecture can be acquired by constructing pattern library applying component reference architecture like development process done. CBD Reference architecture is concern on component, which is supposed to be implemented though analysis and design phase, also possibly apply to entire lifecycle. Figure 6 shows the conceptual process, which are domain analysis, applying design pattern and constructing component-based architecture.



E-Business Agent / Agent Type	System-Level Agent(00)			General Business Activity Agent(10)					Personal Agent (20)	System -Level Agent (30)	Security Agent (40)
	Interoperation Agent (01)	Business Transaction Agent (02)	Workflow Agent (03)	Marketing Agent (11)	Legal Agent (12)	Financial Agent (13)	Negotiation /Contracting Agent (14)	Information Brokering Agent (15)			
Software Agent(SWA)											
Autonomous Agent (AUA)											
Interactive Agent (INA)											
Adaptive Agent (ADA)											
Mobile Agent (MBA)											
Coordinative Agent (COA)											
Intelligent Agent (ITA)											
Wrapper Agent (WRA)											
Middle Agent (MOA)											
Interface Agent (IFA)											
Information Agent (IMA)											
Smart Agent (SMA)											
Hybrid Agent (HYA)											
Heterogeneous Agent (HGA)											
etc											

Fig. 5. Design pattern reference matrix

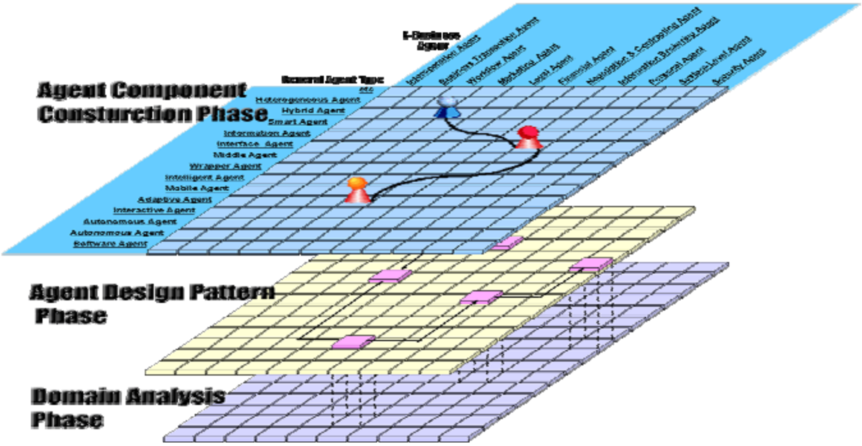


Fig. 6. Agent Development Based Component Reference Architecture

5 Conclusion and Future Works

Agent-oriented technology can help enable the development of e-business agents, which are the next higher level of abstraction in model-based solutions to e-business applications. This technology allows the development of rich and expressive models of an enterprise and lays the foundation for adaptive, reusable business software. Agent-oriented technology can be leveraged to enhance enterprise modeling as well as to offer new techniques for developing applications and infrastructure services.

In this paper, general agent type is classified in 15 categories according to role. e-business agent is classified in 11 categories according to adaptable domain. CBD reference architecture is constructed in 2 dimension based on these categories. In addition, we propose systemical development process based on architecture. This process applies and designs architecture from early domain analysis phase to component development. Design pattern matrix is made in the same architecture mode in component design so that there is a benefit to reduce development time and to have high reusability of design concept. Component reference architecture through agent domain classification is based component development life cycle. Moreover, the development of e-business agent and agent-oriented system can obtain the efficiency through component reuse. In the future work, there needs more study about component integration based CBD reference architecture for e-business agent and agent application system. We also are going to study on the contracting of e-business agent with CBD modeling methodology.

Acknowledgement

"This works supported Korea National Research Foundation (NRF) granted by the Korea Government Scientist of Regional University No.R 2010-0017089".

References

1. Jennings, N.R., Wooldridge, M.: Agent-Oriented Software Engineering. In: Proceeding of IEA/AIE 1999, pp. 4–10 (2009)
2. Odell, James (eds.): Agent Technology, OMG, green paper produced by the OMG Agent Working Group (2010)
3. Papazoglou, M.P.: Agent-Oriented Technology in support of E-Business. Communications of the ACM 44(4), 71–77 (2010)
4. Odell, J., Van Dyke Parunak, H., Bauer, B.: Extending UML for Agents. In: Proceeding Of the Agent-Oriented Information Systems Workshop at the 17th National Conference on Artificial Intelligence (2009)
5. Bauer, B., Müller, J.P., Odell, J.: Agent UML: A Formalism for Specifying Multiagent Interaction. In: Proceeding of 2000 Agent-Oriented Software Engineering, pp. 91–103 (2001)
6. Kim, H.K.: Component Repository and Configuration Management System, ETRI Final Research Report (2000)
7. Nwana, H.S.: Software Agents: An Overview, Software Agent Technologies (1996)

8. Kim, H.K., Han, E.J., Shin, H.J., Kim, C.H.: Component Classification for CBD Repository Construction. In: Proceeding of SNPD 2000, pp. 483–493 (2000)
9. Griss, M.L., Por, G.: Accelerating Development with Agent Components. IEEE Computer 34(5), 37–43 (2001)
10. Brereton, P., Budgen, D.: Component-Based Systems:A Classification of Issues. IEEE Computer 33(11) (2000)
11. Aridor, Y., Lange, D.B.: Agent Design Patterns: Elements of Agent Application Design. In: Proceeding of Autonomous Agents 1998, pp. 108–115 (1998)
12. Heineman, G.T., Councill, W.T.: Component-Based Software Engineering. Addison-Wesley, Reading (2001)
13. Jennings, N.R.: On agent-based software engineering. International Journal of Artificial Intelligence 117(2), 277–296 (2003)
14. Park, K., Kim, J., Park, S.: Goal based agent-oriented software modeling. In: Proceeding of the Seventh Asia-Pacific Software Engineering Conference(APSEC 2000), December 2000, pp. 320–324 (2000)
15. Hara, H., Fujita, S., Sugawara, K.: Reusable Software Components based on an Agent Model. In: Proceeding of 7th International Conference on Parallel and Distributed Systems Workshops, July 2000, pp. 447–452 (2000)

Implementation of Greenhouse Monitoring System Using Ubiquitous Sensor Networks Based on WMN

Jeonghwan Hwang and Hyun Yoe[*]

School of Information and Communication Engineering,
Sunchon National University, Korea
{jhwang,yhyun}@sunchon.ac.kr

Abstract. The USN technology is one of the important technologies to implement the ubiquitous society, which has been used in applications of various fields, and in the agricultural sector, the value-added and productivity are increased by applying it to production management, distribution etc. as well as growing conditions. Particularly, in the greenhouse, the quality and productivity for crops are improved by monitoring and managing growth information for crops and environmental information etc. with the utilization of USN. However, the network topology is frequently changed since sensor nodes building a USN have limited energy resources, and it is not suitable to operate a network by construct a large-scale network through a network extension because packets could be lost during long-distance data transmission. This paper proposes a greenhouse monitoring system applying WMN technologies to a USN to solve problems mentioned above, and implements the system after analyzing the performance through a simulation. In the proposed system, a mesh coordinator is constructed on the data transmission path between sensor nodes and a gateway for monitoring greenhouse environment, and it could be found that this system has higher data transmission efficiency and energy efficiency than the system constructing with conventional USN.

Keywords: Greenhouse, USN, WMN, Agriculture, Ubiquitous.

1 Introduction

The USN(Ubiquitous Sensor Networks) technology is a network system technology that gives computing and network functions to every thing including human living space, appliances, machines etc. to acquire, process and exploit information in real time via networking and communicating among them, which is a important technology to implement the ubiquitous society[1][2], and it has achieved enhancing the standard of human living such as productivity and safety by applying to various industries including distribution, logistics, construction, transportation, defense, and medicine[3][4].

[*] Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 686–698, 2010.

Particularly, in the agricultural sector, productivity is increased and transparent distribution channels are secured for agricultural and livestock products by applying it to production management, distribution etc. as well as cultivation environment, and recently pilot projects and researches are extensively tried to integrate it into various agricultural fields such as greenhouses and livestock barns[5][6][7][8].

In greenhouses, utilizing the USN, environmental change factors of greenhouses including temperature, humidity, illuminance, CO2 and weather etc. are automatically collected to monitor them, and it has created the optimum crop's growing conditions to contribute reducing production costs and improving productivity by managing and controlling crop's growth environment based on crop's growth information and environmental information collected[9][10].

The USN, which is applied to monitor crop's growth information and environmental information in a greenhouse like this, sends data collected from sensor devices to sink nodes or gateways located at a distance in environment without a communication infrastructure, and at this time it is helped by other sensor nodes for long distance data transmission.

During long distance data transmission, since the USN is composed of sensor nodes with limited energy resources, the network topology is frequently changed to increase possibility of packet losses, and there have been arisen problems that the power consumption and overhead of neighborhood sensors delivering data is increased and because the sensor nodes constructing the USN send data by using the broadcast communication.

This paper proposes a greenhouse monitoring system applying WMN(Wireless Sensor Networks) technologies into a USN to solve the presented problems, and constructs mesh networks and implements the system by applying a mesh coordinator corresponding to a relay node between sensor nodes collecting data and a gateway which is final destination node for more efficient greenhouse environmental monitoring after analyzing network performance through a simulation.

This paper is organized as follows. Chapter 2 describes the USN based on WMN for the greenhouse monitoring system, and Chapter 3 compares the IEEE 802.15.4 based sensor networks with the USN based on WMN through a simulation. Chapter 4 explains implementation results of proposed system, finally Chapter 5 completes this paper through a conclusion.

2 USN Based on WMN for Greenhouse Monitoring System

Recently, the IEEE 802.15.4 based USN is widely used for environmental monitoring, and sensor devices sends data to any long distance with other sensor node's help to send collected data to remote sink nodes or gateways in environment without a communication infrastructure.

When sending data to any long distance through a USN, the network topology is frequently changed because the USN is constructed with sensor nodes with limited energy resources, so that possibility of packet loss is increased.

Such a packet loss could become problems for information collection when considering that the USN has a small data size and low data occurrence rate.

In addition, because sensor devices constructing a USN send data via broadcast, there has been arisen problems that power consumption of neighborhood sensors delivering data is increased and the overhead is increased, so eventually data throughput is lowered and the QoS(Quality of service) guarantee becomes uncertain.

To solve problems as mentioned above, the USN based on WMN technology is needed to construct mesh coordinators between sensor device nodes and a gateway which is the final destination node in the IEEE 802.15.4 based sensor networks.

In a multi-hop wireless network, there is a relay node to relay information between two nodes when nodes are communicating, and the communication between two nodes is carried out via the relay node.

The relay node is a mesh coordinator in the WMN, the WMN construction with mesh coordinators improves stability of data transmission via multiple paths, and the load occurred in data communication is concentrated into the coordinator.

The following figure 1 shows an energy efficient sensor network structure using the WMN for the greenhouse monitoring system.

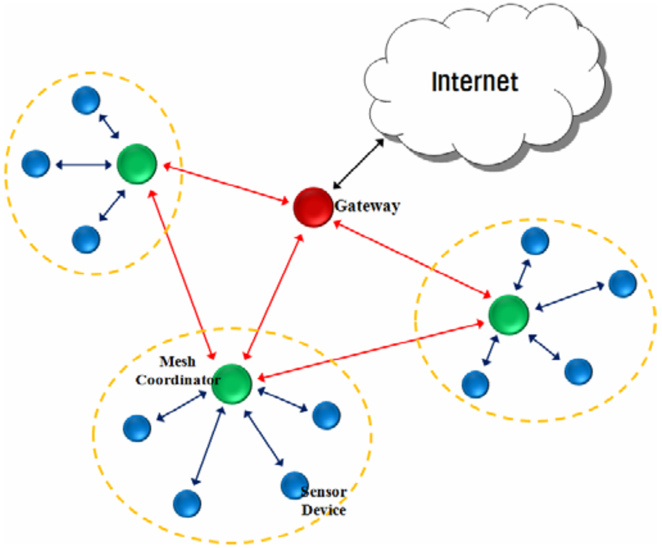


Fig. 1. Energy efficient sensor network structure using the Wireless Mesh Network

Each mesh coordinator constructs the WMN with a gateway as the center, and the mesh coordinator searches the neighborhood sensor devices to associate a connection after completing mesh network configuration and receives the environmental data from sensor devices via point-to-point communication.

At this time, the mesh coordinator sends data received from sensor devices to the gateway via multiple paths and delivers it out.

If a sensor network is constructed with the WMN like this, the transmission distance could be so extended that a large scale network could be constructed through a network extension, and the energy consumption of the relay node and sensor devices could be so decreased that the entire network's life could be lengthened.

3 Simulation and Results Analysis

In order to implement more efficient greenhouse monitoring system, this paper would implement the system applying the WMN technologies to a USN, and a simulation is carried out using the NS-2 as the same conditions as the IEEE 802.15.4 based sensor networks for the performance analysis of network applying the system[11].

In the simulation, data collected into each sensor device is sent to the final gateway via the relay node, which the mesh coordinator is used as the relay node in the sensor network using the WMN, and the sink node is used as the relay node in the sensor network based the IEEE 802.15.4.

To analyze each network's performance, packet transmission rates, delay time and residual energy etc. of the mesh coordinator and the sink node, which correspond the relay nodes of the USN based on WMN and the IEEE 802.15.4 based sensor network, are measured and compared through the simulation.

3.1 Simulation Environment

For the simulation, sensor devices and relay nodes are constructed in an area of 3000 × 3000(m) as the figure 2's topology, which the communication distance between

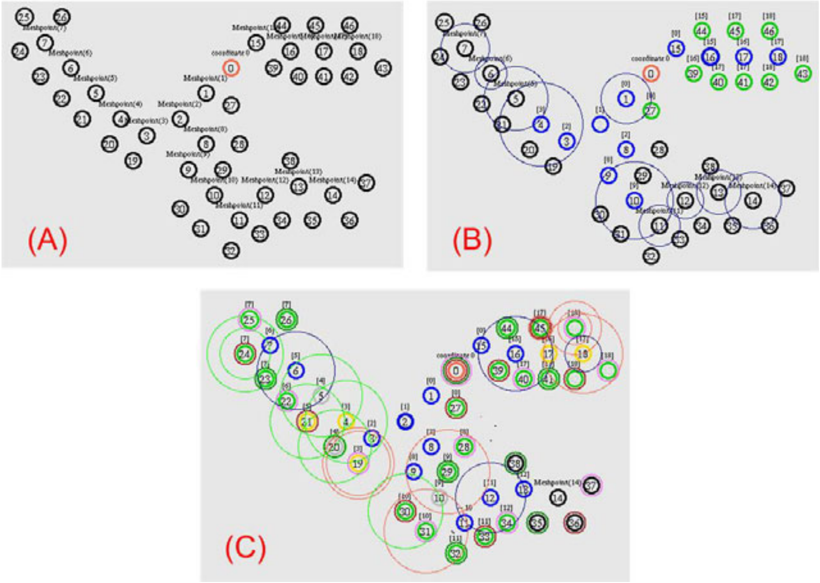


Fig. 2. Simulation network Topology

sensor devices and the mesh coordinator is set as 70 m where the transmit power is 17.4mA, the receive power is 19.7mA, and the PER(Packet Error Rate) is 1% reflecting the experimental result that is the communication radius measured in the actual sensor network[12]. The following table 1 is the environment configured for this paper's simulation.

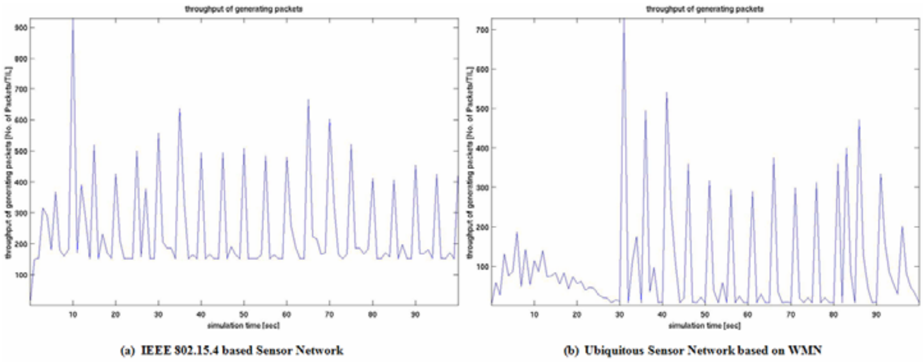
Table 1. Simulation parameter

Configuration options	Configuration values
Simulator	Network Simulator(NS)-2
Channel Type	Wireless Channel
Wireless Propagation Model	Two Ray Ground
Network Interface	IEEE 802.15.4
MAC Type	IEEE 802.15.4
Frequency	2.4 Ghz
Node Distance	70 m
Bandwidth	1 Mbps
Routing Protocol	AODV
Initial Node Energy	3 J
Transmission Power	17.4 mA
Received Power	19.7 mA
Simulation Time	100 Sec
Packet Type	CBR
Packet Size	25 Byte
Sensing Period	5 Sec
Gateway Nodes	1
Mesh Coordinator Nodes	18
Sensor Nodes	28

3.2 Simulation Results and Analysis

The packet throughput, drop packet quantities, and residual energy are measured with time of the mesh coordinator and the sink node corresponding to the relay node as items for results and analysis of the simulation.

The following figure 3 shows the throughput of data packets taking place in each network. It could be found that the packet throughput of the IEEE 802.15.4 based sensor network is higher than the USN based on WMN in general.



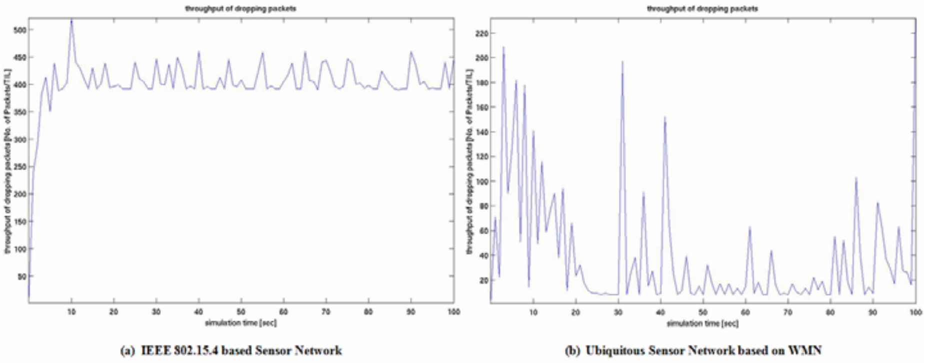
(a) IEEE 802.15.4 based Sensor Network

(b) Ubiquitous Sensor Network based on WMN

Fig. 3. Throughput of generating packets

The following figure 4 shows the drop packet quantities with time in each network. It could be found that the IEEE 802.15.4 based sensor network has high packet quantity dropped from the beginning of initial simulation, and lots of packets are continuously dropped. The reason is that the dropped packets are increased due to frequent packet collisions because the sink nodes, which is the relay node, and the sensor devices use the broadcast communication in the IEEE 802.15.4 based sensor network.

On the contrary, it could be found that the drop packet quantity is high only in the process connecting the mesh network initially between the mesh coordinators, and afterward the drop packet quantity is low at the time when the sensor device sends the environmental data in the USN based on WMN. When the relay node sends the environmental data, the average drop packet quantity for each node is 19 for the sink node, and 14 for the mesh coordinator.



(a) IEEE 802.15.4 based Sensor Network

(b) Ubiquitous Sensor Network based on WMN

Fig. 4. Throughput of dropping packets

The following figure 5 shows the residual energy of the sink node and the mesh coordinator in each network. Since the IEEE 802.15.4 based sensor network uses the broadcast communication, the energy consumption of the sink node is large in general because the data throughput of the sink node and unnecessary operations are increased. On the contrary, since the mesh coordinator of the USN based on WMN uses the point-to-point communication, its energy consumption is smaller than the sink node. The average residual energy of the sink node is 2.14J, and that of the mesh coordinator is 2.59J, so it could be found that the mesh coordinator has higher energy efficiency than the sink node.

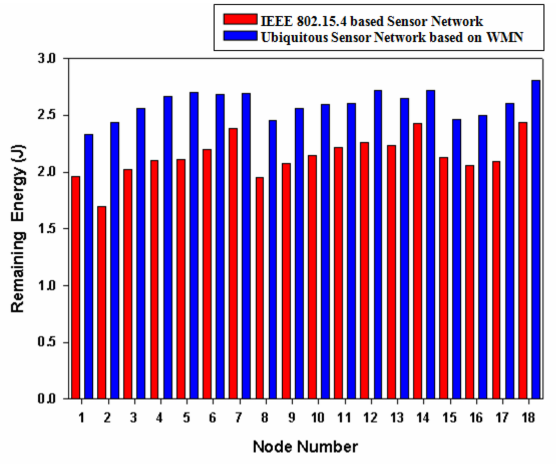


Fig. 5. Average remaining energy of relay nodes

The following figure 6 shows the residual energy of the sensor device in each network. The sensor device of the IEEE 802.15.4 based sensor network has higher energy consumption than that of the USN based on WMN. The average residual energy of the sensor device in the sensor network is 2.52J, and that of the mesh coordinator in the WMN is 2.74J, so it could be found that the mesh coordinator has higher energy efficiency than the sink node.

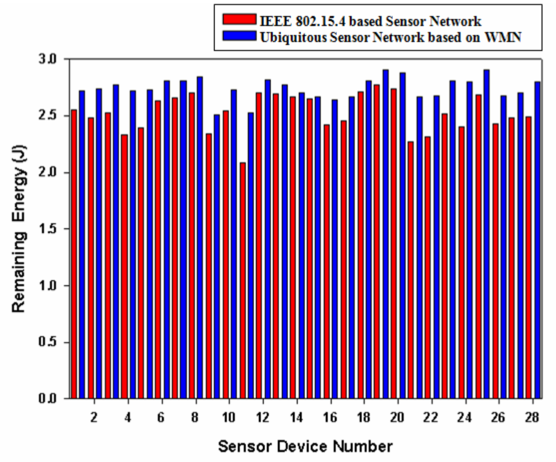


Fig. 6. Average remaining energy of sensor devices

4 Implementation of Proposed Greenhouse Monitoring System

It could be found that the USN based on WMN has better performance than the IEEE 802.15.4 based sensor network in terms of the data transmission efficiency and the energy efficiency through the simulation results, and this paper implements the system applying the USN based on WMN technology to a greenhouse for efficient greenhouse monitoring. The following figure 7 is the block diagram of greenhouse monitoring system applying the proposed USN based on WMN technology.

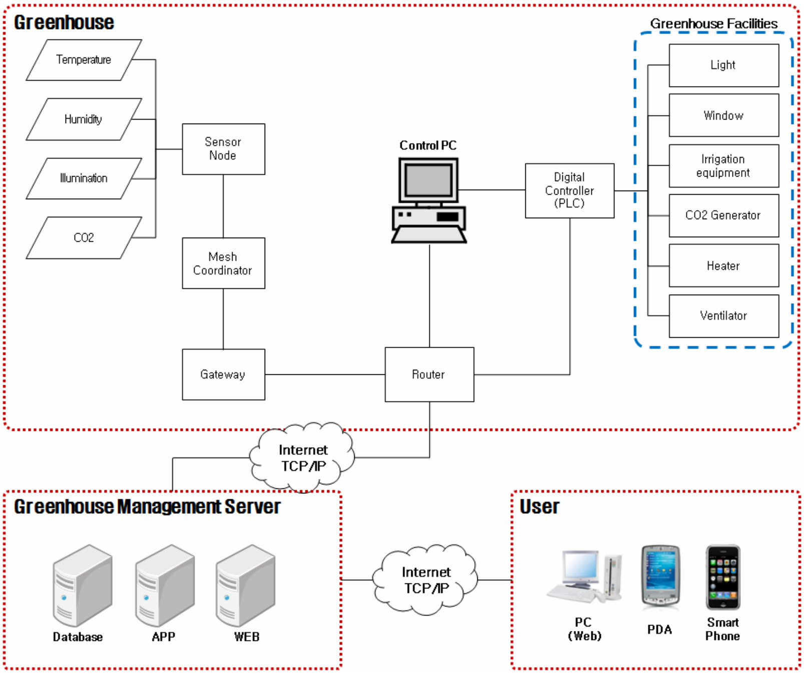


Fig. 7. Block diagram of greenhouse monitoring system applying the proposed USN based on WMN technology

4.1 System Structure

As the figure 8, the proposed greenhouse monitoring system is composed of the physical layer, which is composed of sensors, CCTV, and greenhouse facility, the middle layer, which supports communication between the physical layer and the application layer and maintains greenhouse's growing conditions as the optimum status by making greenhouse's information as a database to support monitoring and control services, and the application layer, which has interfaces to support the greenhouse environmental monitoring and facility control service.

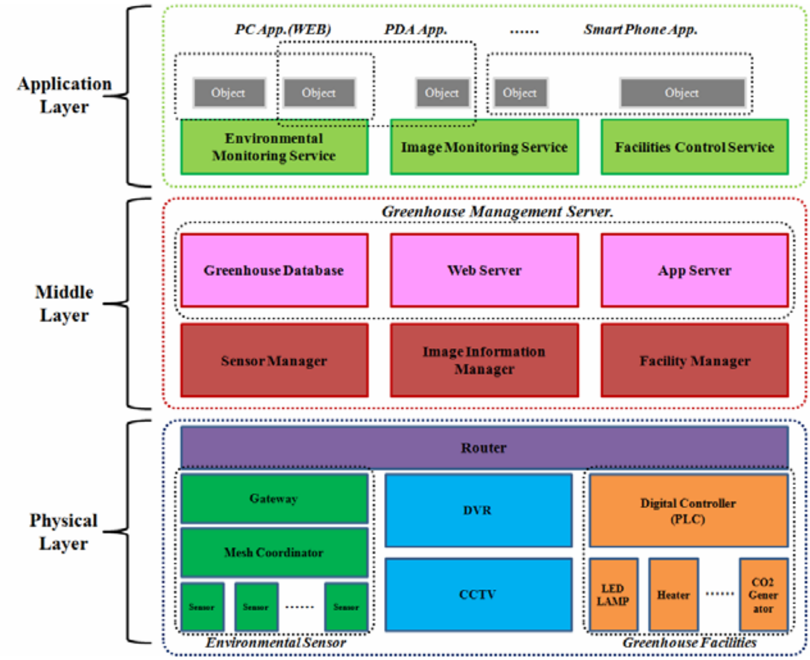


Fig. 8. Proposed greenhouse monitoring system structure

The physical layer is composed of sensors, which collect greenhouse's environmental information, CCTVs, which collect image information of greenhouse, and the greenhouse facility, which makes the optimum growing conditions of greenhouse.

The sensors measure the environmental information such as illuminance, temperature, humidity, wind direction, wind speed, EC, pH, CO2 etc. affecting crop's growth inside/outside the greenhouse, and send the collected information to the gateway via the mesh coordinator.

The CCTVs are installed inside/outside the greenhouse, the internal CCTV is installed to collect image information and crop's image information, and the external CCTV is installed to prevent risks such as burglaries and fires etc.

The greenhouse facility is composed of the environmental control facilities including lighting, ventilator, hot-air blower, CO2 generator to control the greenhouse environment affecting the crop's growth such as illuminance, temperature, EC, pH, CO2 etc., and is controlled through a PLC.

The middle layer is composed of a sensor manager, which manages the environmental information collected from sensors of the physical layer, a image information manager, which manages the image information collected from CCTVs, a facility manager, which manages the greenhouse facility, a database, which stores the greenhouse information, and a greenhouse management server, which monitors the greenhouse and controls the greenhouse facility.

The sensor manager converts the environmental information inside/outside the greenhouse collected from physical layer's sensors into the form to store in the greenhouse database, converts into the unit suitable to measured elements, and stores the converted data into the greenhouse database using the update queries.

The facility manager receives control signals from the greenhouse management server to operate or manage the greenhouse facility through the PLC, and stores conditions, operating time and control times of such greenhouse facility into the greenhouse database.

The image information manager sends images taking from the CCTVs to the greenhouse management server to provide stream data into the Web, and classifies the image into the greenhouse ID and camera number etc. to store into the database.

The greenhouse database is in charge of storing the greenhouse environmental data including illuminance, temperature, humidity, CO2 etc. collected from sensors installed inside/outside the greenhouse, the image data collected from CCTVs, the environmental control facility's conditions and operating time/control times, the environmental reference values for automatic control and informing conditions into each table.

The greenhouse management server is located between the user and the greenhouse database, informs the greenhouse environmental data stored in the greenhouse database to the user at regular intervals, and controls the corresponding greenhouse facility automatically or provides the alarm service via the Web and SMS etc. comparing the greenhouse facility control table with the environmental reference value stored in the condition notice table.

The application layer is composed of application services to support various platforms such as the Web, PDA, and smart phones etc., which could provide users the greenhouse monitoring service and the greenhouse control service.

4.2 Implementation

Sensors are installed inside/outside the greenhouse as the figure 9 to measure the environmental information inside/outside the greenhouse including illuminance, temperature, humidity, wind direction, wind speed, EC, pH, CO2 etc. in order to measure the environmental information inside/outside the greenhouse, and the collected information is sent to the gateway via the mesh coordinator.

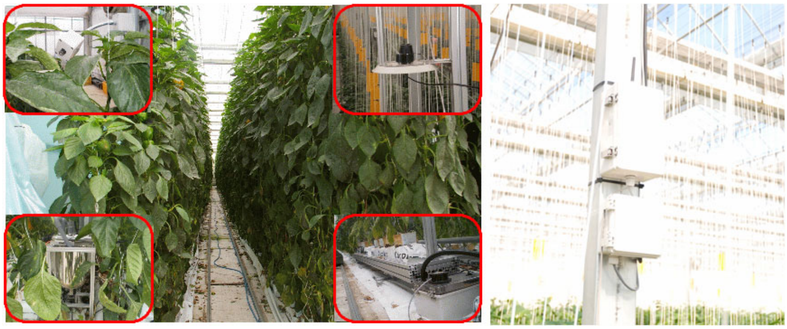


Fig. 9. Environmental Sensor Device and Mesh Coordinator

In order to make the optimum greenhouse growth conditions based on the greenhouse environmental information collected from sensors installed inside/outside the greenhouse, the environmental control facility is installed in the greenhouse as the figure, and the PLC controller is installed as the figure 10 to control them.

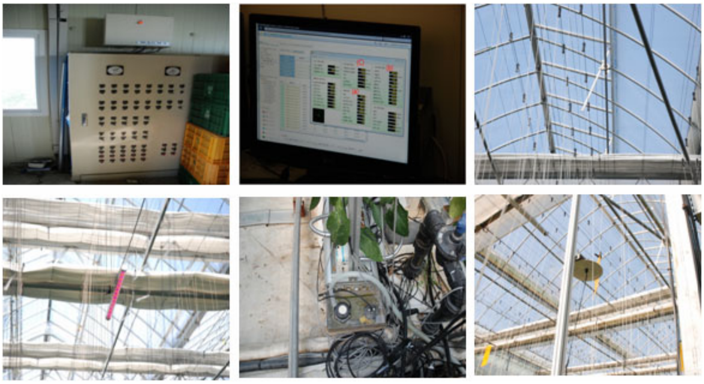
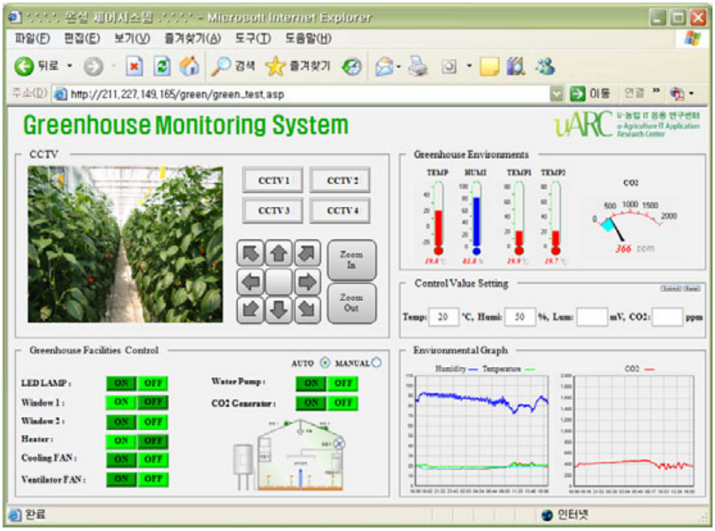


Fig. 10. PLC controller and Greenhouse Facilities

The GUI for the administrator is developed as the Web environment as the figure 11 to monitor and control the greenhouse, the Tomcat-6.0.20 is used as the WAS, and the Mysql version 5.0, which is the most stable version among currently released versions, is used as the database.



http://211.227.149.165/green/green_test.asp

Temp: 20 °C, Humi: 50 %, Lum: mV, CO2: ppm

AUTO ◉ MANUAL ○

LED LAMP :	ON OFF	Water Pump :	ON OFF
Window 1 :	ON OFF	CO2 Generator :	ON OFF
Window 2 :	ON OFF		
Heater :	ON OFF		
Cooling FAN :	ON OFF		
Ventilator FAN :	ON OFF		

Fig. 11. Greenhouse Monitoring System WEB GUI

As a result of applying the greenhouse monitoring system proposed as mentioned above into the actual greenhouse, the environmental and image information of greenhouse could be collected through sensors and CCTVs, monitor and control the conditions of greenhouse through the GUI, and obtain the greenhouse environmental graph as the following figure 12.

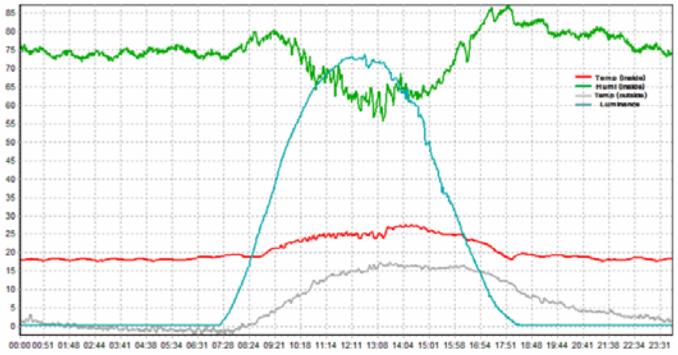


Fig. 12. Paprika Greenhouse Environment Data Graph

5 Conclusions

This paper proposed the greenhouse monitoring system applying the WMN technologies into a USN for more efficient greenhouse monitoring, and the system was implemented after analyzing the network's performance through a simulation.

As a result of the simulation, the sensor network with the WMN has lower packet throughput and fewer packets dropped in the data transmitting process comparing to the IEEE 802.15.4 based sensor network, and it could be found that the energy efficiency is high due to low power consumption of sensor devices.

The greenhouse environmental monitoring system was implemented in the actual greenhouse by constructing the mesh coordinator on data transmission paths between sensor nodes and a gateway based on such results, and the growth environment of greenhouse could be monitored and controlled via the GUI as a result of implementation.

It could be expected several advantages that the environmental data collected from sensor devices in the greenhouse could be sent to the long distance through the network extension, and the working life of nodes could be lengthened because the power consumption of nodes is minimized.

Acknowledgements. This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2010-(C1090-1021-0009)).

References

1. Akyildiz, I.F., et al.: A survey on Sensor Networks. IEEE Communications Magazine 40(8) (2002)
2. Chong, C.-Y., Kumar, S.P., Hamilton, B.A.: Sensor networks: evolution, opportunities, and challenges. Proc. IEEE 91(8), 1247–1256 (2003)
3. Misic, J., Shafi, J., Misic, V.B.: Avoiding the bottlenecks in the MAC layer in 802.15.4 low rate WPAN. In: Proc. of ICPADS, pp. 363–367 (2005)
4. Pyo, C.-S., Chea, J.-S.: Next-generation RFID / USN technology development prospects. Korea Information and Communication Society. Information and communication 7-13, 7–13 (2007)
5. Jeong, B.-m.: Foreign u-Farm Service Model casebook, Korea National Information Society Agency, NCA V–RER-06005
6. Kwon, O.-b., Kim, J.-h.: A Basic Direction for Building Agricultural Radio Frequency Identification Logistics Information System, Korea Rural Economics Institute, M85, 12 (2007)
7. RFID Journal, http://www.rfidjournal.com/article/articleview/2229/1/1/
8. Korea Agricultural Trade Information(KATI), The state of sweet pepper industry in korea, Korea Agro-Fisheries Trade Corporation (2009)
9. Lee, M.-h., Shin, C.-s., Jo, Y.-y., Yoe, H.: Implementation of Green House Integrated Management System in Ubiquitous Agricultural Environments. Journal of KIISE 27(6), 21–26 (2009)
10. Park, D.-H., Kang, B.-J., Cho, K.-R., Sin, C.-S., Cho, S.-E., Park, J.-W., Yang, W.-M.: A Study on Greenhouse Automatic Control System Based on Wireless Sensor Network. Wireless Pers. Commun. (2009)
11. Fall, K., Varadhan, K.: The ns Manual (formerly ns Notes and Documentation). The VINT Project, A Collaboration between researches at UC Berkeley, LBL, USC/ISI and Xerox PARC
12. Kim, H.-S., Jung, W.-S., Yun, C.-Y., Oh, Y.-H.: A Research on a Cross Layer Protocol with Communication Radius in Wireless Sensor Networks. Journal of Korea Information and Communication Society 33(4), 113–123 (2008)

A Study of the Energy Efficient MAC Protocol Apply to Ubiquitous Livestock Farming

Hochul Lee, Jeonghwan Hwang, and Hyun Yoe*

School of Information and Communication Engineering,
Sunchon National University, Korea
{hclee,jhwang,yhyun}@sunchon.ac.kr

Abstract. The ubiquitous pig farm monitoring system is designed so that it can be applied to modern agriculture and attached MAC Protocol that the Link Quality based Transmit Power Control MAC(LPMAC) protocol, which is becoming bigger and more advanced. In the immediate previous research, a large scale ubiquitous pig farm monitoring system was designed and it was applied to actual pig farms. It was possible to realize the system, where the raising efficiency of pig was maximized by installing the temperature/humidity sensors and video camera, transmitting the measured information to caretaker in real time. When the measured environmental information values detected by the sensor would get out of optimum raising environment of pigs, the system transmitted the content to the caretaker. Simultaneously, the environment control facilities such as humidifier, air conditioner and window opener/closer were activated in order to maintain the environmental information at uniform and pleasant values. And, it was found that the LPMAC protocol is proper as the MAC to be applied to the livestock farming.

Keywords: Ubiquitous, WSN, Monitoring, Pig Farm, Energy Efficient MAC.

1 Introduction

Human wishes for more convenient and safe living are being realized in the improvement of all fields such as transportation, distribution, agriculture, health, medical science and defense industry together with the rapid advance of IT technology[1]. Such efforts bring in ubiquitous world into the reality in combination with u-IT technology and it further pursues the ultimate u-Green IT by combining with the topic of environment-friendliness.

This research designed and realized the 'ubiquitous pig farm monitoring system', which can be a model in the 'pig farm management and monitoring field' during the realization of u-GreenIT. Most of existing pig farm management systems just monitor the temperature/humidity and control the temperature. The caretaker should stay nearby pig farm, visually check the pig farm and take care of control and monitoring of pig farm. These make the system difficult to be applied to the Korean cattle-raising industry now[2]. The ubiquitous pig farm monitoring system suggested in this research

* Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 699–707, 2010.

has complemented such shortcomings. It can monitor the environmental information of pig farm and control the pig farm facilities from remote location by controlling the temperature, humidity and bad smell intensity in the pig farm.

This research paper consists as following. In chapter 2, related works and MAC protocols. In Chapter 3, the ubiquitous pig farm monitoring system using sensors has been realized and tested. In Chapter 4, the results are summarized as the conclusion. It is expected that this ubiquitous pig farm monitoring system will optimize pig farm operation efficiency in the future and enhance the productivity of cattle and quality of livestock products.

2 Related Works

2.1 Ubiquitous Stall Monitoring System Using IP-USN

This research was carried out as a part of joint technology development's support business supported by the Small and Medium Business Administration. This was researched from June 2009 to May 2010 and is a stall monitoring and control management system using IP-USN technology. And this is a system that monitors a stall environment such as temperature, humidity and ammonia, etc. by installing sensors for environmental measurement at a stall, and that when abnormalities such as an abnormal environment and fire, invasion and theft, etc. occurred, it informs the abnormalities to a producer so that he can cope with it quickly. In addition, the scalability capable of monitoring large-scale stalls is provided by using sensor nodes applied with IP technology, and application models suitable for various rural environments were constructed by guaranteeing mobility of sensor nodes. WSN is also used as USN in Korea[3].

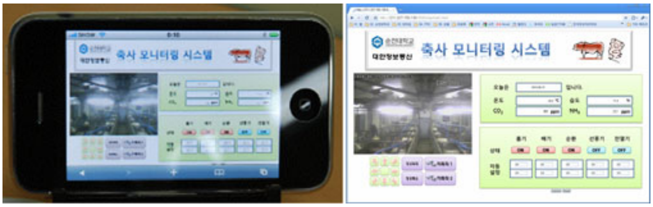
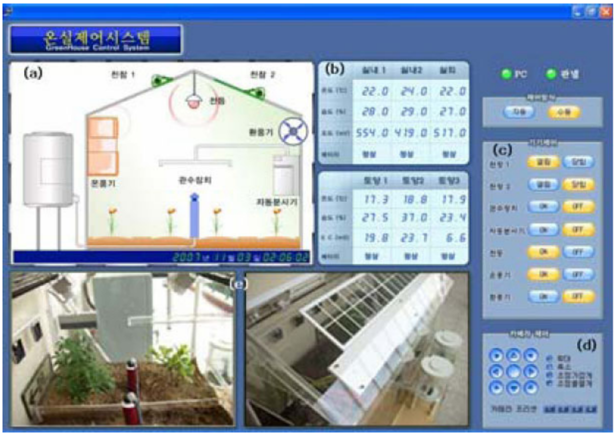


Fig. 1. u-Stall Monitoring System GUI (Smart phone and WEB)

2.2 Green House Integrated Management System

In order to create a ubiquitous agricultural environment as well as measure temperature, humidity, flux of solar radiation, carbon dioxide, ammonia, wind velocity and rainfall, etc. influencing cultivation environments, the present research composed sensor networks in greenhouses and developed a greenhouse environment's monitoring system that controls a device having influence on change of environmental factors such as ventilation fans, windows, heating, humidification and illumination, etc. This system can grasp a greenhouse state in real time through the Internet, and can perform remote monitoring of a state inside a greenhouse through CCTV with the naked eye, and is a system that can manage a greenhouse at any place where its remote control is possible by delivering a warning signal through a personal terminal, even if a manager doesn't watch the system in case of generation of hindrance or an abnormal condition[4].



(b)	실내1	실내2	실외
온도 (℃)	22.0	24.0	22.0
습도 (%)	28.0	29.0	27.0
조도 (lux)	554.0	419.0	517.0
[illegible]	정상	정상	정상

	토양1	토양2	토양3
온도 (℃)	17.3	18.8	17.9
습도 (%)	27.5	37.0	23.4
E.C (mS)	19.8	23.1	6.6
[illegible]	정상	정상	정상

Fig. 2. Green House Integrated Management System GUI

2.3 LPMAC Protocol

LPMAC(Link quality based on Power control MAC) protocol transmits isochronous packet using maximum power in the synchronization process before it sends the data to the neighborhood node. The nodes which received this packet measures the RSSI and LQI, determines the most optimum power and exchanges the maximum power value during synchronization process.

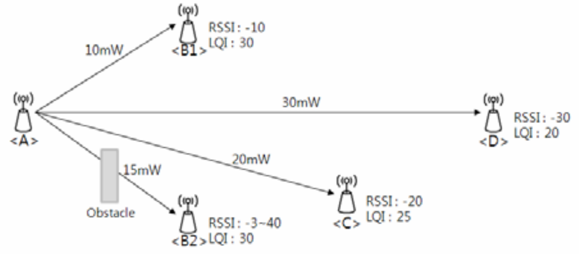


Fig. 3. LPMAC concept diagram

The smaller the RSSI of received signal, the bigger power should be suggested; however, it examines the communication quality at the time, and, if the communication quality is good, it suggests the transmission power one level lower. The transmitted power selects the power with good communication quality by considering the communication quality with relevant node every time transmitting/receiving of the data would be made with neighborhood node and SYNC packet. This process is done in the synchronization process regardless of transmitting/receiving nodes. The energy consumption can be reduced more by such control of transmitting power[5].

2.4 Sensor Node

The sensor node developed by Sunchon National University ITRC Research Center will be applied to this research. This sensor node can collect the information of leaf wetness, leaf temperature, greenhouse temperature/humidity and control the relay by one sensor. MSP430 MCU is applied to the CPU and CC2420 RF module of Chipcon Co. is used as the data transmission/reception device. The MSP430 microprocessor has 16 bit RISC structure and it works in very fast speed with its 48 Kbyte program memory and 10Kb RAM. 3.6V battery is used for power supply [6]. The MAC protocol for sensor nodes is LPMAC protocol.

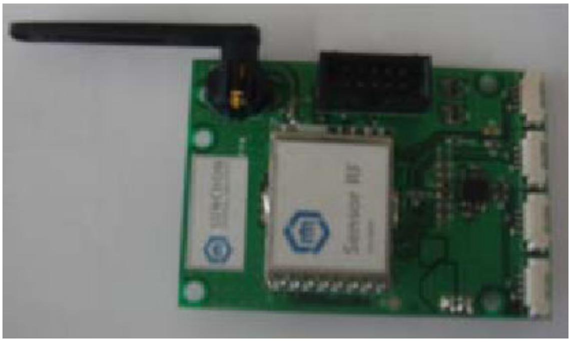


Fig. 4. u-ARC was developed Sensor node module

3 Paper Preparation

3.1 Architecture of Pig Farm

The sensor node installed in pig farm is introduced as chapter 2.4. These WSN sensors will be installed in the pig farm. These sensors make the wireless network together with environmental control devices, including the WSN sensor gateway in the pig farm. These sensors detect the temperature, humidity, illumination intensity and bad-smell intensity. They send the sensor-measured values to pig farm managing device in a certain fixed cycle. The sensor node is sealed in a plastic box because of the box will protect the sensor node. The system can keep the pleasant raising environment inside pig farm by controlling the environmental control devices connected to the sensors and operating the pig farm operating devices.

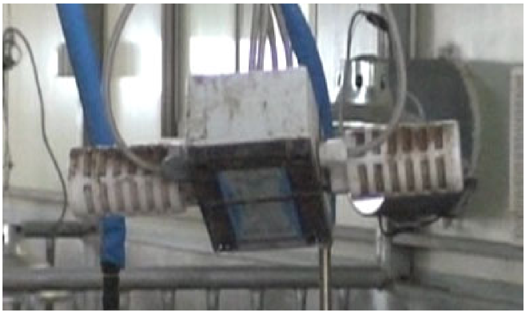


Fig. 5. Plastic box to protect the sensor node

Fig. 6. Environment Control Device

IP cameras were installed to monitor the inside of pig farm 24 hours by video. This camera monitors and records the pig farm inside 24 hours. It is used to find out the causes of accidents such as theft or accident, in addition to real-time monitoring of pig farm status. The recorded videos are sent to pig farm management server. There, they are classified by pig farm ID and camera ID before they are stored in the database.

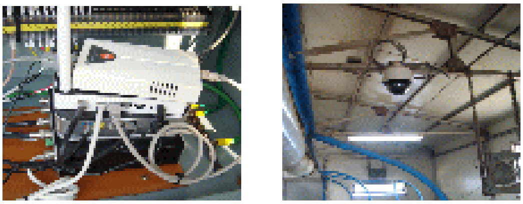


Fig. 7. Video Server and IP-Camera

3.2 Sensor Node Installation and Network Topology

The sensor nodes will be installed at every 2m along the pig pence lined up in reference to the pig farm. They will be installed alternately for the pig tail direction wall and the door of pig pence. Installation will continue to the culture media with 2m distance in reference to the culture media with sensors installed. The overall location shape of sensor nodes is grid-type with 2m distance. The sink node to transmit collected data to the server will be located in the center of 4 * 50 grids. The shape of sensor nodes location is grid shape; however, the network topology is a star topology in reference to the sink node in the center of the grid.

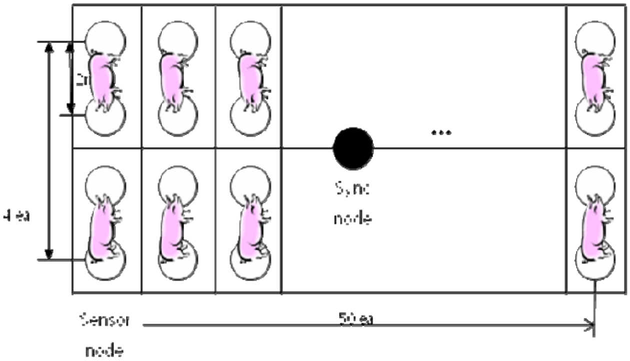


Fig. 8. Sensor location and Network topology

3.3 Pig Farm Management Server

The environmental measurement data in stream type from pig farm is ‘parsed’ by interpreter before they are stored in the database. Simultaneously, they are sent to the pig farm caretaker so that he/she can know the environment change in real time.

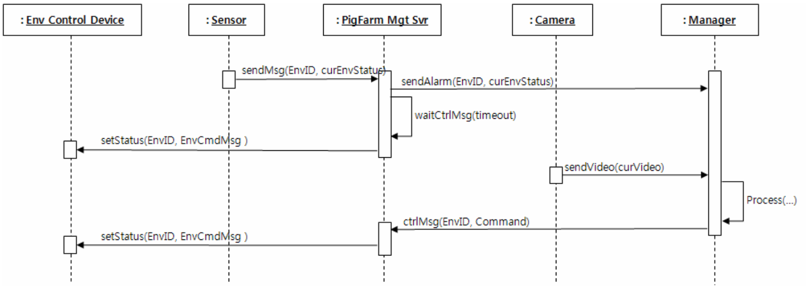


Fig. 9. Sequence Diagram of Message exchange[7]

There are two ways to keep the environmental status in the pig farm. The first is automatically control the environment from system based on the designated environment standard data. The second is directly control the system as the caretaker requires. At this time, the system remembers the control pattern of the caretaker and system feeds the pattern back to the environmental standard data and applies it to future automatic environment control mode.

With these execution results, it was possible to verify that the environmental information and video information inside pig farm are collected through sensors and cameras and the pig farm can be monitored and controlled through real-time GUI. The contents expressed in the GUI are overall configuration of pig farm and its operation status. The latest measured environmental information is shown on the screen. The measurement is done in 5 minutes cycle and it can be also seen as a graph by querying the previous status change. The result of apply LPMAC, was submitted good activation system with legacy MAC Protocol.

With these implementation results, it was possible to verify that the environmental information and video information inside pig farm are collected through sensors and video surveillance cameras and the pig farm can be monitored and controlled through user-intuitive GUI.

3.4 Measuring Environment

In order to complete the design and evaluate the performance of this system, two test-beds were established in two pig farms located in Suncheon of Chollanam-do. Pig farm A had sensors and WSN. Pig farm B had sensors, WSN and additional environment controlling devices. Two pig farms were located in the same area with 5m distance from each other. The most optimum raising environment for pigs is 15℃～18℃ temperature and 60～80% humidity; therefore, the environment controlling devices operate when the condition would escape from these ranges.

3.5 Result and Analisys

The environment sensors installed at pig farm send the measured data to server every 5 minutes. Measurement period was from 00:00 hours of September 26th, 2010 until 24:00 hours of September 30th. The server applies the measured data to the decision making mechanism of pig farm environment control and controls the pig farm environment.

The measurement result of natural state pig farm A showed that the temperature and humidity change substantially. This test was done during the season change from spring to summer, when the climate change was rather big. The cattle are supposed to receive severe stress in this kind of big temperature change. Figure is the temperature and humidity change in pig farm, which has ubiquitous pig farm monitoring system installed and operating. It does not have big change in temperature and humidity compared to pig farm A.

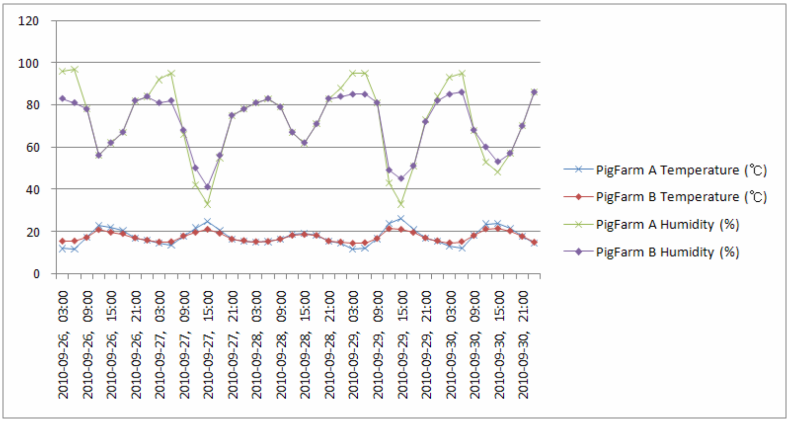


Fig. 10. Environmental changes of Pig Farm

After the installation of this system, estimated data based on the measured data was generated. The resultant figures of changed temperature and humidity inside pig farm after the operation of inside control devices agreed with the thermometer and hygrometer installed in the pig farm. The pig farm showed less change rate in temperature and humidity compared to the pig farm without separate control. This measurement result suggests that the pig farm operating by suggested ubiquitous pig farm monitoring system and it attached of LPAMC, that is more effective than the pig farm operation in natural state.

4 Conclusion

This research suggested ubiquitous pig farm monitoring system which can effectively manage the raising environment of pigs by collecting the environment information such as temperature, humidity and illumination intensity and automatically, or, manually control the various devices in the pig farm.

Sensor nodes and various devices were installed in the actual pig farms to verify the performance of suggested system. The test gave the result that the pig farm status is accurately monitored and the devices are correctly controlled. Through these, it was possible to confirm that the suggested system can maintain the raising environment of pig at the most optimum condition by confirming the pig farm inside status through GUI and pig farm video system from remote location using various sensors and devices.

The IT age is passing and ubiquitous age is coming; however, primary industries like agriculture are not getting the benefit of informatization age. Pig-raising industry is getting bigger and becoming corporations. However, without systematic and efficient pig-raising control, it is difficult to expect high productivity and high quality pork production. The ubiquitous pig farm monitoring system will contribute in the saving of labor force in the pig-raising farmers and in the production of high quality pork. It will further contribute in the securing of competitiveness by pig-raising industry, by way of joining pig-raising industry with ubiquitous technology, called 'u-IT'.

Acknowledgment

This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2010-(C1090-1021-0009)).

References

1. Mainwaring, A., Polastre, J., Szewczyk, R., Culler, D., Anderson, J.: Wireless Sensor Networks for Habitat Monitoring. In: International Workshop on Wireless Sensor Networks and Applications (2002)
2. Hwang, J.-h., Kang, H.-j., Lee, M.-h., Shin, C.-s., Yeo, H.: RFID/USN based livestock-disease spreading early detection system. In: KICS Autumn Conferences (2008)

3. Yoe, H., Loe, J.-m., Jeon, G.-h., Lee, M.-h., Hwang, J.-h., Lee, J.-w., Yoon, D.-h., Yeo, I.-j., Kim, H.-k.: Final Report: Ubiquitous Stall Monitoring System using IP-USN. Small and Medium Business Administration, Korea (2010)
4. Lee, M.-h., Shin, C.-s., Jo, Y.-y., Yoe, H.: Implementation of Green House Integrated Management System in Ubiquitous Agricultural Environments. Journal of KIISE 27(6), 21–26 (2009)
5. Lee, H.-c., Hwang, J.-h., Lee, M.-h., Kim, H.-k., Yoe, H.: A Power Control Scheme for an Energy-Efficent MAC Protocol. In: Kim, T.-h., Adeli, H. (eds.) AST/UCMA/ISA/ACN. LNCS, vol. 6059, pp. 586–596. Springer, Heidelberg (2010)
6. Park, D.-H., Kang, B.-J., Cho, K.-R., Sin, C.-S., Cho, S.-E., Park, J.-W., Yang, W.-M.: A Study on Greenhouse Automatic Control System Based on Wireless Sensor Network. Wireless Pers. Commun. (2009)
7. Lee, H.-c., Ju, H.-d., Hwang, J.-h., Yang, C.-j., Yoe, H.: Design and Implementation of Ubiquitous Pig Stable Monitoring System. In: KICS Autumn Conference (2009)

Design and Implementation of Middleware for GreenHouse Based on Ubiquitous Sensor Network

Ji-woong Lee, Jeong-hwan Hwang, and Hyun Yoe*

School of Information and Communication Engineering,
Sunchon National University, Korea
{leejiwoong,jhwang,yhyun}@sunchon.ac.kr

Abstract. The USN middleware technology is used to filter lots of duplicate data collected from many sensor networks and convert the raw data into meaningful information for users to send it to applications, and provides services to make users could decide contextual information quickly and correctly through the data mining technique and analysis method. Even though it has been presently carried out the studies on such a USN middleware to apply it for various fields such as administration, medicine, science, transportation, and logistics etc., there are very few studies on the middleware suitable to agricultural environment which applications of IT technology have not been sufficient relatively comparing to other industries. In particular, for controlled agriculture, there are many difficulties on user's decision-making for efficient crop production due to a number of environmental factors affecting crop production. In order to solve such problems, this paper is trying to propose a USN middleware suitable to agricultural environment, which could collect greenhouse's environmental information and optimally manage crops through facility automation. The proposed middleware is composed of a sensor manager, context manager and control manager, which collects a variety of data from heterogeneous sensor networks, processes the collected data into information suitable to user's demand, and sends it to controllers of controlled agriculture, so that it could support users to be provided various application services and make decisions adequate to conditions.

Keywords: USN, MiddleWare.

1 Introduction

The recent trend in research and development on ubiquitous computing technologies is towards the direction to provide users optimum services suitable to conditions through context awareness, inference and cooperation based on data collected from various sensor nodes[1]. The USN technology is one of the ubiquitous computing's core technologies, which is applied to various fields including production, distribution, logistics, medicine, welfare, environment etc. to pursue enforcement of human life's convenience, improvement of quality of life, promotion of welfare, and

* Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 708–718, 2010.

security[2]. To build such the ubiquitous application services easily, a middleware is needed to connect the RFID/USN's hardware with the applications or the enterprise systems[2]. The middleware is a technology to filter lots of data collected from many heterogeneous RFID/sensor devices, process the event data, and then abstract it into meaningful information[3], and to send and process a great number of contexts and data arisen in the ubiquitous environment more efficiently[4]. Even though researches on the USN middleware are currently in progress for various fields, there are very few researches on the middleware focused on application services in agricultural environment that the application of IT technology is insufficient relatively comparing to other industries[5]. In particular, for the controlled agriculture, the production and the quality of crops is affected by the consistent management of various environmental factors such as temperature, humidity, insolation, CO2, ammonia, wind speed, rainfall etc. affecting crop's growth, and the precision control of environmental control devices including ventilator, windows, heater, lighting, image processor etc., so many difficulties are arisen in producer's decision making. This paper would like to propose an USN middleware suitable to agricultural environment, which could collect greenhouse's environmental information and manage crops optimally through facility automation in order to solve problems in such controlled agriculture environment. The proposed middleware is designed to collect and monitor the environmental information from sensors installed in the facility, and to provide the optimum service to the agricultural application service system by controlling the facility control devices through the corresponding context information processing when a problem is arisen, which helps user could be provided various application services and make a decision suitable to the situation. This paper is organized as follows. Chap. 2 explains the related researches, Chap. 3 analyzes requirements on the middleware to design the middleware based on the results, Chap. 4 implements the designed middleware, and finally Chap. 5 draws the conclusion of this paper.

2 Related Works

2.1 A design of Context Aware Middleware Based on Web Service in Ubiquitous Environment

Context-aware technologies for ubiquitous computing are necessary to study the representation of gathered context-information appropriately, the understanding of user's intention using context-information, and the offer of pertinent services for users. [6] this paper propose the WS-Cam(Web Services based Context-Aware Middleware) framework for context-aware computing. WS-CAM provides ample power of expression and inference mechanisms to various context-information using an ontology-based context model. this also consider that WS-CAM is the middleware-independent structure to adopt web services with characteristic of loosely coupling as a matter of communication of context-information. this paper describe a scenario for lecture services based on the ubiquitous computing to verify the utilization of WS-CAM. this paper also show an example of middleware-independent system expansion to display the merits of web-based services. WS-CAM for lecture services represented context-information itodomaits as OWL-based ontology model effectively, and

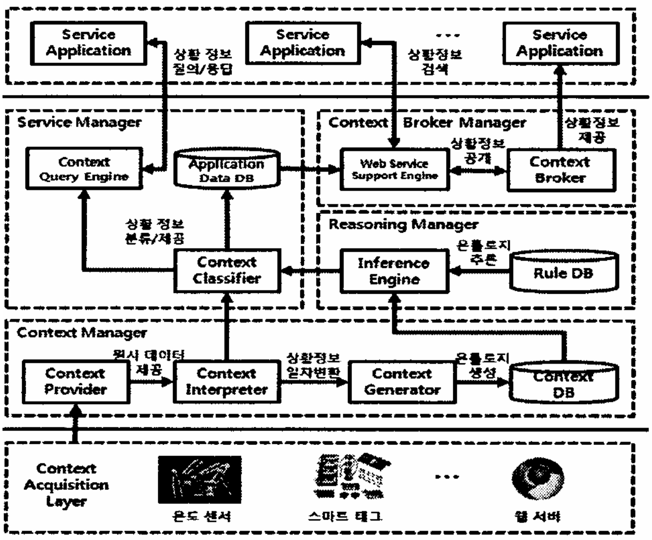


Fig. 1. The data flow diagram of Ws-CAM Framework

confirmed the information is inferred to high level context-information by user-defined rules. this paper also confirmed the context-information is transferred to application services middleware-independently using various web methods provided by web services[6].

2.2 Implementation of an Application System Using Middleware and Context Server for Handling Context-Awareness

Context-awareness is a technology to facilitate information acquisition and execution bysupporting interoperability between users and devices based on users' context. It is one of the most important technologies in ubiquitous computing. this paper propose a middleware and a context server for dealing with context-awareness in ubiquitous computing and implement an application system using them.[7]

The middleware proposed in this work plays an important role in recognizing a moving node with mobility by using a Bluetooth wireless communication technology as well as in executing an appropriate execution module according to the context acquired from a context server.[7]

In addition, the proposed context server functions as a manager that efficiently stores into a database server context information, such as user's current status, physical environment, and resources of a computing system. [7]

Finally, this application system implemented in this work one which provides a music playing service based on context information, and it verifies the usefulness of both the middleware and the context server developed in this work.[7]

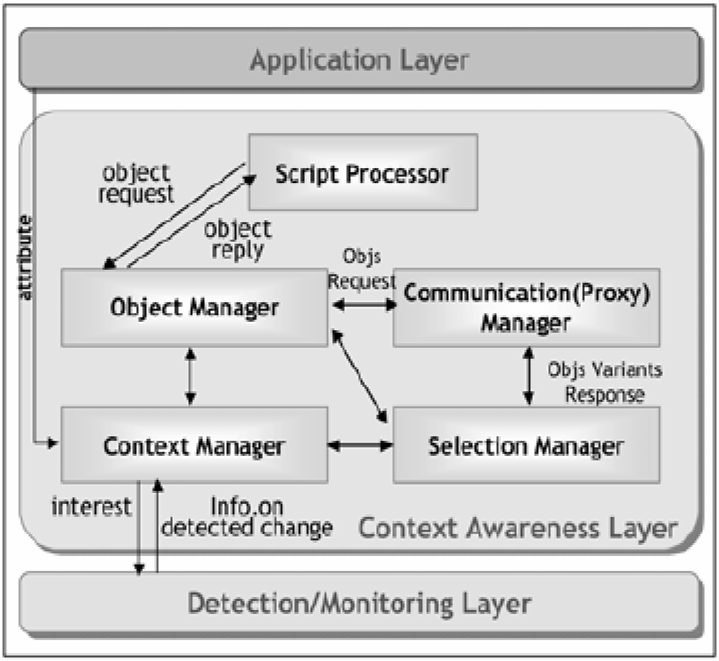


Fig. 2. Middleware Structure

3 MiddleWare

3.1 Middleware Requirement

To control and manage an agricultural facility efficiently, it should be considered the environmental aspect of system and the function of USN middleware. First, in the environmental aspect of system, when the difference of temperature between crops and the air is more than 4˚C in the facility, the condensation is taken place, so crops are damaged due to disease, and the production is significantly different depending on the light environment, temperature and humidity environment in the facility. In addition, since producers may suffer a loss due to unnecessary heating bills, in order to cope actively with it[8][9], it is installed the sensors for environmental information including temperature, illuminance, humidity, CO2 etc., and the control system such as heater, CO2 controller, wind speed/wind direction controller, ventilator etc. for the optimum environment. Second, the fundamental functions of middleware are the multiple query processing of collected services, management of sensing and meta information, creation of context information for sensing information, intelligent event processing required from the application layer[2]. Among them, sensors collect environmental variables (temperature, humidity, illuminance etc.) to provide adequate services for agricultural facilities, event processing is carried out to process data for pre-registered conditions if the certain condition is satisfied, and the collected environmental information is compared and analyzed with existing collected data. In addition, a service is provided users to make adequate decisions by creating contextual information through prediction and inference.

The middleware is designed on the basis of such requirements.

3.2 Middleware Design

Fig. (3) shows the structure of USN middleware proposed in this paper for controlled agriculture automation, which is composed of a sensor manager (SM), context manager (CM), and control manager (CTM). The SM has a function to collect information taken place in the facility and to take charge of communication between middlewares, the CM has a function to analyze the raw data collected by the SM to convert it into actually usable information and to store and manage information. The CTM controls and manages the facility's device based on information analyzed by the CM.

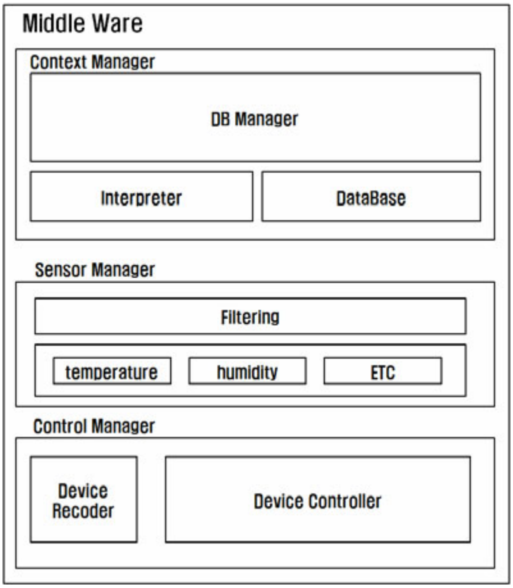


Fig. 3. USN Middleware Structure

3.2.1 Sensor Manager

The sensor manager is a module to deliver environmental information, which takes charge of interfaces between physical sensors and computers. The sensor manager carries out a function to collect information from the sensors including temperature, humidity, illuminance, CO2 etc. installed in the facility and the control devices such as ventilator, heater etc., sends constant clock signals to synchronize sensors in order to correctly transfer data between sensors and gateways without errors, and removes duplicate data by filtering to send it to the CM since the efficiency is decreased due to lots of data when receiving duplicate data.

3.2.2 Context Manager

The context manager could effectively manage the various contextual information to intelligently provide it to users. Such contextual information may be collected from the various sensors installed in the facility, and also collected via the Web such as information of the other facilities or the surrounding area[6]. The context manager takes charge of managing functions to acquire, process, represent, and store information for users and surrounding environment of users obtained from the various sources as above[6]. The context manager is composed of a context interpreter and a context DB manager as the Fig. (4).

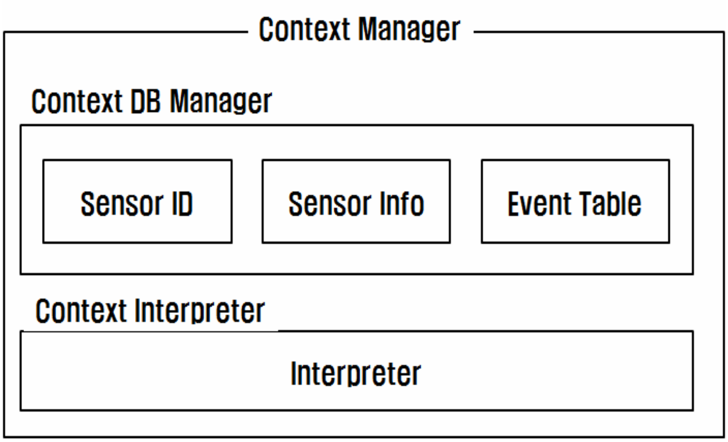


Fig. 4. Context Manager Structure

The context interpreter takes charge of converting the raw data collected from sensors into semantics that could be comprehended at the user level. Such converted information is stored in the database through the context DB manager. The context DB manager is comprised of Sensor ID, SensorInfo, and EventTable[10].

Table 1. Sensor ID

Sensor ID	Sensor function
1	Humidity
2	Co2
3	Fan
…	…

As the Table 1, Sensor ID table is comprised of sensor's ID and sensor's function attribute, and SensorInfo table allocates periods, time, measured values to be collected from sensors and assigns Group ID for each role of sensors.

Table 2. Sensor Info

Sensor ID	Location	Sample Cycle	Time	Value	Group ID
1	3-2 …	250	201009100423	40	3
2	4-1	250	201009100423	47	2
3	5-4	250	201009100423	30	1

The contextual information could be created through the data analysis module, which analyzes conditions of environment in the facility and crops based on such stored information, and the data mining technique, and the table is constructed as the Table 2 for intelligent event processing required by users. Certain problems occurring in the facility, i.e. many problems that temperature/humidity is too high to cultivate crops or the concentration of CO2 becomes so high that it has a bad influence on photosynthesis of crops, are predefined in the event table as the Table 3, and the facility is controlled if the problem is arisen. The interaction of CM is as follows.

Table 3. Event Table

Event	Group Id average Value
Turn on the Fan	40
Turn off the Fan	27
Turn on the light	30
…	…

The Group ID is given according to the sensor's function as the Table 2, the average value of sensor information values collected for each group is stored in the database as the form identical to the Table 4. The facility is automatically controlled if the event condition is satisfied on the basis of this value.

Table 4. Group ID

Group ID	Average Value
1	17
2	22
3	25
…	…

3.2.3 Control Manager

The control manager is composed of a device controller and a device recorder. The device controller requests the contextual information to control, and the device recorder records the current condition of control device to refer for the next service request.

Exploiting these two functions, the CTM uses the contextual information received by the CM to adequately control various control devices at the locations where the event is arisen, and sends the information to users in case of emergency. scenario for operating.

3.3 Scenario for Operating

Fig. 5 is the process of entire system. This SM periodically collects the environmental information such as temperature, humidity, CO2, wind direction, wind speed etc. from the sensor network installed in the agricultural facility. The collected information is stored in the database through filtering in order to remove duplicate data. The CM creates the contextual information from the stored information through the data mining and analysis, and uses the contextual information to send the control signal to the CTM through the predefined event manager. The CTM controls the controller in the corresponding area to efficiently operate based on the received control signal.

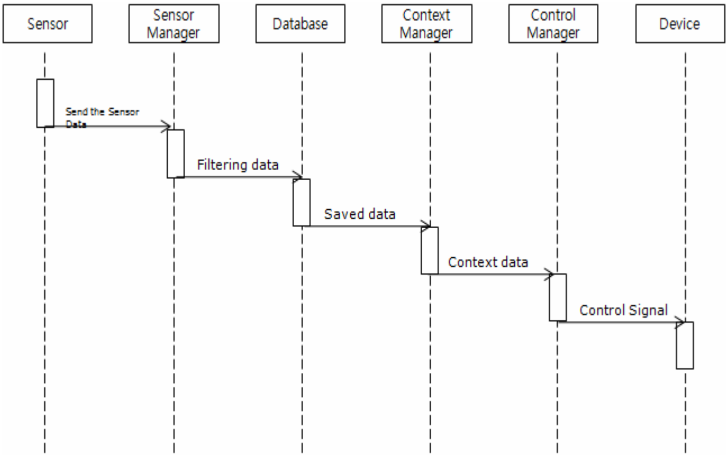
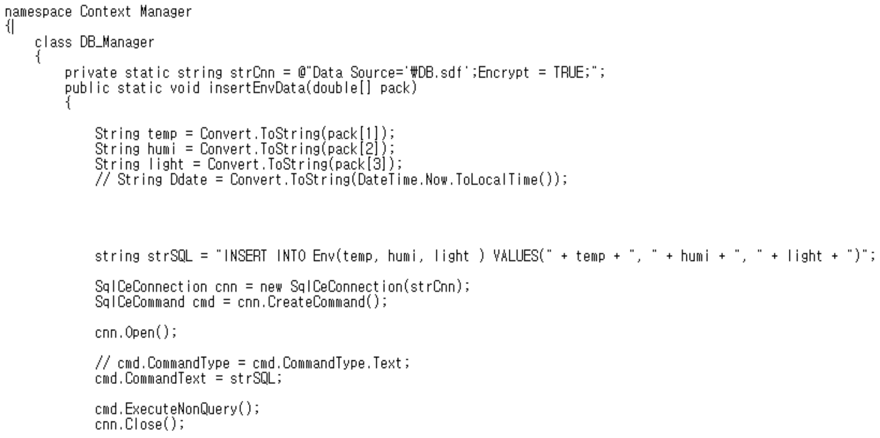


Fig. 5. Scenario Flowchart

4 Implementation

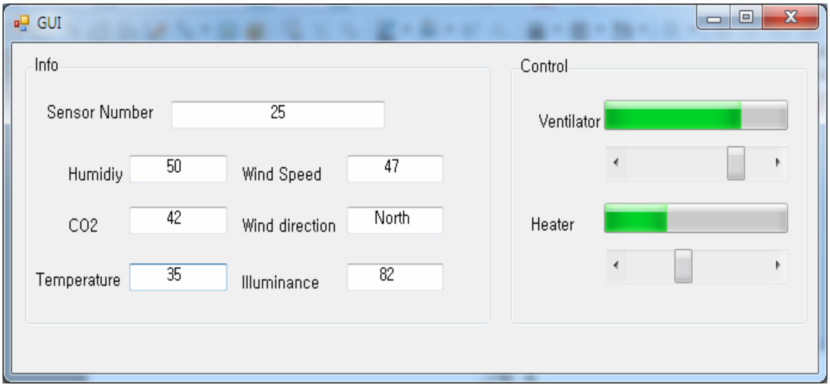
The proposed middleware is aimed at implementing the middleware for the Greenhouse suitable to the agricultural environment, and data collected through the sensor network is experimented for the event extraction according to the given conditions for the aim. In addition, it is constructed to confirm the results through the GUI implemented by the Microsoft Visual Studio 2005 C#.

Fig. 6 is the CM implemented with the C#, which is part of codes storing the environmental information received from the SM into the database, and Fig. 7 is the GUI to confirm the results of the proposed middleware. Through the GUI in the facility of Fig. 7 (Info), it could be confirmed the environmental information values such as temperature, humidity, illuminance, CO2, wind direction etc. collected from sensors, the intelligent event processing is confirmed through the event notification window as Fig. 8 opened when the contextual information exceeds the reference value, and the performance of middleware is confirmed by controlling various devices such as ventilator, heater etc in the facility through the Fig. 7 (Control).



```
namespace Context Manager
{|
    class DB_Manager
    {
        private static string strCnn = @"Data Source='₩DB.sdf';Encrypt = TRUE;";
        public static void insertEnvData(double[] pack)
        {

            String temp = Convert.ToString(pack[1]);
            String humi = Convert.ToString(pack[2]);
            String light = Convert.ToString(pack[3]);
            // String Ddate = Convert.ToString(DateTime.Now.ToLocalTime());



            string strSQL = "INSERT INTO Env(temp, humi, light ) VALUES(" + temp + ", " + humi + ", " + light + ")";

            SqlCeConnection cnn = new SqlCeConnection(strCnn);
            SqlCeCommand cmd = cnn.CreateCommand();

            cnn.Open();

            // cmd.CommandType = cmd.CommandType.Text;
            cmd.CommandText = strSQL;

            cmd.ExecuteNonQuery();
            cnn.Close();
```

Fig. 6. Part of Context Manager Code



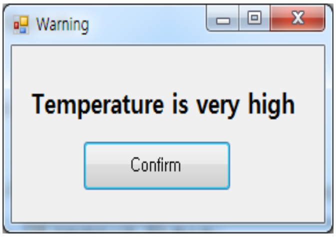
Sensor Number 25

Humidiy 50 Wind Speed 47

CO2 42 Wind direction North

Temperature 35 Illuminance 82

Fig. 7. GUI



Temperature is very high

Fig. 8. Event Notification

5 Conclusion

This paper designs and implements the middleware to control the facility according to the contextual information collected from sensors for the agricultural facility automation suitable to the agricultural environment. The middleware is composed of the sensor manager, context manager, control manager, which the sensor manager sends various environmental information to the context manager, the context manager creates the contextual information and analyzes the agricultural environment based on the event, and the agricultural facility is controlled through the control manager, so it is minimized the problems that could be arisen in the facility. In addition, it is designed to monitor the information collected from sensors to support decision-making in the agriculture site. It is expected that the high profit would be given to the farm if the stability and reliability of facility is secured and the collected growth condition of crops is exploited through the middleware proposed in this paper.

Acknowledgements. This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2010-(C1090-1021-0009)).

References

1. Ju, H.-D., Im, H.-J., Lee, M.-H., Yoe, H.: Design of middleware in WSN for large scale glasshouse. In: Proceedings of the Korean Institute of Maritime Information and Communication Sciences Conference, pp. 351–353 (2007)
2. Kung*, S.H., Kang*, Y.H., Yoo, J.H.: USN Based Middleware Software Design for Agriculture and Stockbreeding. In: Proceedings of the KAIS Fall Conference, pp. 788–791 (2009)
3. Hwang, J.G., Cheong, T.S., Kim, Y.I., Lee, Y.J.: ETRI, Trends of RFID Middleware Technology and its Aplications. Electronics and telecommunications trends 20(3) (93) (2005)
4. Lee, K.-j., Song, S.K., Youn, H.Y.: A New Context-Oriented Middleware for supporting Exact Context-Awareness in Ubiquitous Environment. In: Korea Computer Congress, vol. 33(1(D))
5. Kung, S.H.: The Design of Fungus Cultivating System based on USN. Korean Institute of Information Technology, 34–41 (2007)
6. Song, Y.-R., Woo, Y.-S.: A Design of Context-Aware Middleware based on web Services in Ubiquitous Environment. The Korea Institute of signal Processing and Systems 10(4), 225–232, 1229–9480
7. Shim, C.-B., Tae, B.-S., Chang, J.-W., Kim, J.-K., Park, S.-M.: Implementation of an Application System using Middleware and Context Server for Handling Context-Awareness, vol. 12(1), pp. 31–42 (2006)
8. Lee, J.-w., Lee, H.-c., Hwang, J.-h., Cho, Y., Shin, C., Yoe, H.: Design and Implementation of wireless sensor networks based paprika Green house system. Communications in Computer and Information Science 78, 638–646 (2010)

9. Jeong, W.-j., Lee, J.h., Kim, H.c., Bae, J.H.: Dry Matter Production, Distribution and Yield of Sweet Peper Grown under Glasshouse and Plastic Greenhouse in Korea. Journal of Bio-Environment Control 18(3), 258–265 (2009)
10. Park, H.-C., Lee, J.-s., Jang, K.-W., Lee, J.-W., Park, J.-H., Kang, S.-Y.: Network platform for integrated information exchange on ship. In: Proceedings of the ISME 2009 (2009)

A Context-Aware Streaming Agent Method for Optimal Seamless Service in a Vertical Handover Environment

Inwhee Joe* and Jong-Yong Park

Department of Electronics and Computer Engineering, Hanyang University,
Seoul, 133-791 South Korea
iwjoe@hanyang.ac.kr

Abstract. We propose a novel context-aware streaming agent to achieve optimal high quality seamless streaming service in heterogeneous wireless networks and mobile terminals and to utilize high WLAN bandwidth in mobile phones. The major difference between our proposed method and existing methods is the use of mobile terminal oriented context-aware streaming adaptation. We outline a streaming network architecture among the streaming server, gateway, and mobile terminal, an extended RTSP protocol and a streaming quality level selection algorithm in the mobile terminal. We consider vertical handover between HSDPA and 802.11g/n and take advantage of H.264 Scalable Video Coding (SVC). Our new method makes optimal seamless streaming service possible for various network situations, terminal capabilities, and user preferences. Streaming agents first determine adequate quality level based on user preference, terminal constraints, dynamic network conditions and battery status. Then a gateway requests an adequate SVC layer to be sent to the server based on the agent's requested quality level. Agent and gateway communicate using an extended RTSP protocol which enables content quality modification during streaming. Our novel streaming agent method improves user satisfaction. By selecting an "optimal network condition" policy, we can improve content and maximize the potential of the 3.28 SVC layer. By minimizing power consumption, we can reduce power consumption by 13% by processing a different SVC layer.

Keywords: Streaming agent, Seamless streaming service, Heterogeneous wireless networks, Mobile phones.

1 Introduction

Recently produced mobile terminal equipped 3G and WLAN modules simultaneously and LCD more then WVGA size and codec which decoding capability

* This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2010-(C1090-1021-0009)).

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 719–726, 2010.

more then H.264 Main Profile. Some products can encode HD 720p. We focused on improving streaming service in enhanced wireless networks and mobile terminals. This paper describes a novel context-aware streaming agent that uses a streaming network, extended RTSP protocol, and a streaming agent in its mobile terminal. We utilize QoS and mobility for vertical handover and H.264 SVC.

2 Related Studies

This paper considers vertical handovers, end-to-end QoS[4], H.264 SVC codec[5], multimedia protocols (RTSP, RTP, and SDP), streaming servers, gateways, agents and, terminal capability. Adaptive streaming, and streaming using H.264 SVC, are areas of active research. MPEG-21 DIA and JVT's H.264 SVC are already standardized, and release 6 of the 3rd Generation Partnership Project (3GPP) added adaptive streaming in PPS (Packet-Switched Streaming). Streaming adaptation in MPEG-21 DIA [6], 3GPP PSS and H.264 SVC are based on client feedback. With these approaches, quality is initially fixed and it is impossible to dynamically change content quality factors such as frame rate and resolution. 3GPP PSS adaptation techniques [2] employ bit-stream thinning and switching, meaning that only frame rate change is possible. In this paper, we create an agent in a mobile terminal that can adapt content quality and select access networks for maximum user satisfaction.

3 Context-Aware Streaming Agent

This paper references established standards and on-going standardization efforts. We exclude detailed discussion of QoS and mobility in vertical handovers and quality evaluations of the proposed method. We focus on streaming agents, network entity, protocol and use-case.

3.1 Streaming Network

We propose a streaming network using H.264 SVC. Fig 1 depicts the streaming server, streaming gateway, and mobile terminal.

streaming server has H.264 SVC encoded contents. Each dataset is input in two files. One is encoded with a scalable baseline profile for mobile environments, and the other is encoded with a scalable high profile for high quality content. Each encoded H.264 SVC file has three layers. The Base Layer is for horizontal/vertical handover in poor transmission conditions. Enhanced Layer 1 is used under most conditions. Enhanced Layer 2 is used under particularly good transmission conditions. The scalable baseline profile uses only temporal scalability in a mobile environment. It requires more decoding power and is easy to adapt in mobile environments. The scalable high profile uses spatial scalability in this context. Additional temporal/SNR scalability can be used. The two profiles have a total of six quality levels [Table 1].

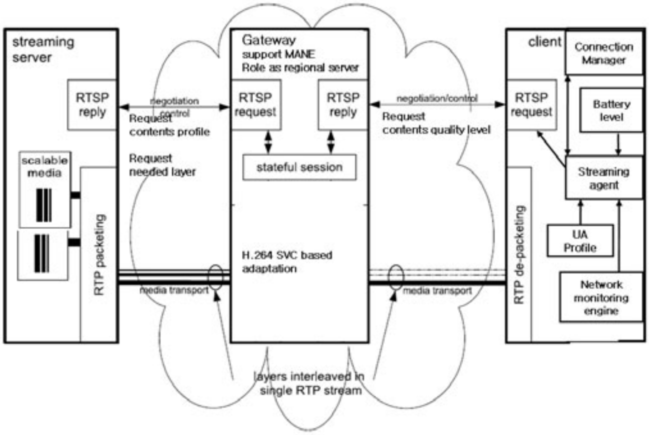


Fig. 1. Streaming network model

Table 1. Sample content format stored in the streaming server

Content Profile	Scalable Baseline Profile			Scalable High Profile		
Layer	BL	EL1	EL2	BL	EL1	EL2
Resolution	240p	240p	240p	480p	720p	1080p
Frame /second	10	20	30	30	30	30
Bitrates (bps)	180 - 250K	250 - 512K	512 - 800K	1 - 1.5M	2 - 3M	4 - 6M
Quality level	0	1	2	3	4	5

Content type does not depend on whether HSDPA or WLAN accesses the network. The mobile terminal's role is to select an appropriate quality level, while the server's and gateway's role is to send contents or layer to the mobile terminal. The server can send only enough layers of content to avoid IP backbone network traffic congestion for the uni-cast stream, and sends a whole stream of appropriate profile contents to the gateway for multi-cast and broadcast streams to support heterogeneous networks and devices. The streaming gateway [1] is located between the access core network and the IP backbone network. The access core network manages 3G and WLAN access networks simultaneously. It also should support Media Aware Network Elements (MANE) and truncate the SVC stream, which is received from the server, up to the needed layer depending on the access network's bandwidth and mobile terminal's request. The gateway should also support the assembly of multiple RTP streams from the server to a single RTP stream for the terminal because wireless environments have bandwidth constraints. Only needed layers are assembled and sent to the terminal.

A streaming agent sends the terminal's UA Profile and selected quality level to the gateway at first negotiation. The gateway requests and negotiates with the server depending on the terminal's request. During the horizontal/vertical handover, signal quality and battery status change, and the agent determines the best quality level based on user policy and terminal constraints, then sends the selected quality level to the gateway. When the terminal moves from 3G to WLAN, signal quality is improved, and the agent requests increased content quality. When the battery level is low, the agent requests a decreased content quality level for more battery time. Changing the quality level is limited by user policy, codec and LCD constraints.

3.2 Extended RTSP Protocol

To support the terminal agent, the RTSP protocol between the terminal and gateway and between the gateway and server must be changed. Extending the RTSP based on 3GPP PSS [3] is for dynamic contents quality change to support agent's request. The extended RTSP protocol is applied to the terminal and gateway. This paper describes an extended RTSP protocol between the terminal and gateway but can also be applied between the gateway and server.

The existing RTSP is modified in the DESCRIBE command and the response to the support agent request in the first negotiation.

1) c→s DESCRIBE : Send UA Profile to gateway with additional terminal information and selected contents.
2) s→c Reply of DESCRIBE : Gateway sends SDP to terminal. SDP contains gateway (or server) supported content quality level. In this method, the gateway can decrease the quality level by monitoring the network.

Next, RTSP is extended to support the dynamic content quality level issued by the terminal when adding the QoS_CHANGE command. Gateway and terminal can be issued at anytime during the streaming.

1) c→s / s→c QoS_CHANGE : Terminal requests a new content quality level to reflect the wireless access network's signal status and battery status.
2) s→c / c→s Reply of QoS_CHANGE : Gateway can accept by sending "OK" or sending supportable contents to the server or gateway.

3.3 Context-Aware Streaming Agent

The streaming agent selects the optimal quality level based on the constraints of the twork and terminal. [Table 2] lists constraints. The agent has information about user preference and terminal constraints. Information on network status and battery level is available from the connection manager and the power manager. The agent is able to change the streaming content's quality level dynamically during streaming service. Figure 2 shows how to select quality level.

Table 2. Constraint items of streaming service

Terminal Constraints and User Policy (Fixed Constraints)	Constraint of Air Conditions and Battery (Dynamic Constraints)
User policy for vertical handover	Vertical handover case
Maximum supported H.264 SVC codec	Horizontal handover case
LCD resolution	Good signal quality
User policy for battery and network throughput	High signal quality
	Battery status

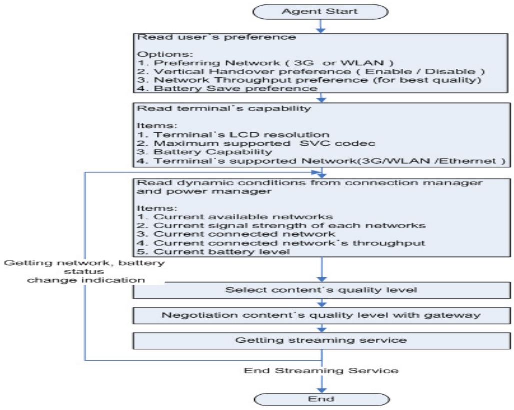

Agent Start
Read user`s preference
Options:
1. Preferring Network (3G or WLAN)
2. Vertical Handover preference (Enable / Disable)
3. Network Throughput preference (for best quality)
4. Battery Save preference
Read terminal`s capability
Items:
1. Terminal`s LCD resolution
2. Maximum supported SVC codec
3. Battery Capability
4. Terminal`s supported Network(3G/WLAN /Ethernet)
Read dynamic conditions from connection manager and power manager
Items:
1. Current available networks
2. Current signal strength of each networks
3. Current connected network
4. Current connected network`s throughput
5. Current battery level
Getting network, battery status change indication
Select content`s quality level
Negotiation content`s quality level with gateway
Getting streaming service
End Streaming Service
End

Fig. 2. Quality level selection process

The streaming agent selects quality level based on user preference. There are four options.

1) Preferred Network
Select a user's preferred network. The selection of the initial network connection for streaming service is related to network cost.
2) Vertical Handover Preference
By enabling this option, a user can obtain quality service from various networks during vertical handover.
3) Network Throughput Preference
User obtains maximum quality with maximum throughput. Battery consumption will increase.
4) Battery Save Preference
Power is saved at the cost of quality.

4 Experimental Results and Discussion

In this experiment, we obtained data by testing Palma-CE1-Conditions using SVC reference software [7]. This software extracts the SVC layer using bitrates at various levels of video resolution and frame rate. We identified seven layers as listed in Table 3 by layer number, video resolution and frame rates. CIF 15fps and 4CIF 30fps have bitrates with different SNR scalability to compensate for bitrate gaps between layers. Layer 0 is QCIF 15fps at a maximum of 128Kbps. 128Kbps is guaranteed in most 3G services by APN for streaming service QoS. From layer 0 to layer 3, we assume that each layer is a layer of the SVC file encoded with a Scalable Base Profile. From layer 4 to layer 6, we assume that each is a layer of the SVC file encoded with a Scalable High Profile. Layer 0 and layer 4 are the base layers of each SVC file and are compatible with H.264 AVC. Table 3 also shows the content quality ratio and decoding power consumption ratio. These two factors are assumed to measure content quality and decode power consumption. They are relative values compared with layer 6 (4CIF, 60fps, 2018Kbps). Table 4 lists testing terminals. Fig 3 shows available bandwidth variation in the test case. At the 50 minute point, the line is split into two lines. The upper line is for WLAN after vertical handover. The lower line is for 3G without vertical handover to WLAN.

Table 3. Conditions for each layer

Layer (Quality Level)	Resolution	Frame Rate (fps)	Bitrates (Kbps)	Contents Quality Ratio	Decoding Power Consumption Ratio
0	QCIF	15	128	10%	10%
1	CIF	15	256	20%	20%
2	CIF	15	384	30%	30%
3	CIF	30	512	50%	50%
4	4CIF	30	1024	70%	70%
5	4CIF	30	1536	80%	80%
6	4CIF	60	2018	100%	100%

Table 4. Test terminals

Layer (Quality Level)	Resolution	Frame Rate (fps)	Bitrates (Kbps)	Contents Quality Ratio	Decoding Power Consumption Ratio
A0	CIF	CIF30	X	-	X
B0	4CIF	4CIF60	X	-	X
C0	4CIF	4CIF60	X	-	X
A1	CIF	CIF30	O	Network	O
B1	4CIF	4CIF60	O	Network	O
A2	CIF	CIF30	O	Battery	O
B2	4CIF	4CIF60	O	Battery	O

By using a PC simulation program that implements streaming agent methods, we obtain the selected quality level. Table 4 and Table 5 show test results.

In Table 5, A0, A1, A2 can obtain CIF 30fps content. The streaming agent with A1 and A2 installed obtained 2.51 more layers than an agent with only A0 installed. B0 and B1 provide enough terminal capability to obtain 4CIF 60fps

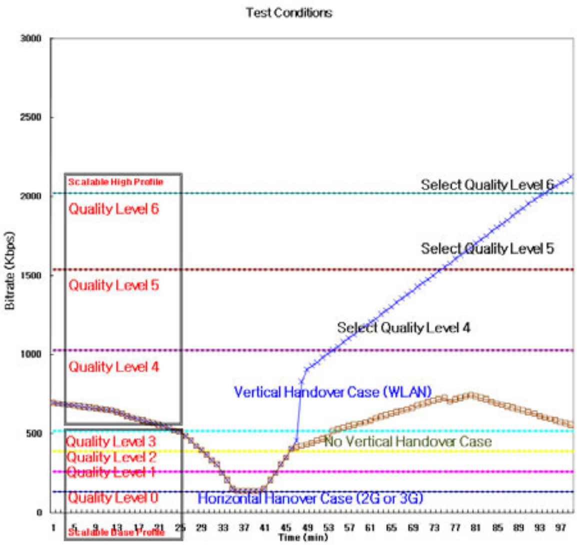


Fig. 3. Test scenario: variation of available bandwidth

Table 5. Test Result I

Section	1	2	3	4	5	6	7	8	9	10
Time (min)	1-25	26-33	31-44	34-44	45	46-48	49-54	55-75	76-94	95-100
Duration	25	5	3	11	1	3	6	21	19	6
3G only case	3	2	1	0	1	2	2	3	3	3
3G/ WLAN handover case	3	2	1	0	1	2	3	4	5	6
Selected Contents Layer										
A0	0	0	0	0	0	0	0	0	0	0
B0	0	0	0	0	0	0	0	0	0	0
C0	3	0	0	0	0	0	3	3	3	3
A1	3	2	1	0	1	2	3	3	3	3
B1	3	2	1	0	1	2	3	4	5	6
A2	3	2	1	0	1	2	3	3	3	3
B2	3	2	1	0	1	2	3	3	4	4

Table 6. Test Result II

	LCD Resolution	Supported Codec	Agent installed	Agent's policy	Vertical Handover Support
A0	CIF	CIF30	X	-	X
B0	4CIF	4CIF60	X	-	X
C0	4CIF	4CIF60	X	-	X
A1	CIF	CIF30	O	Network	O
B1	4CIF	4CIF60	O	Network	O
A2	CIF	CIF30	O	Battery	O
B2	4CIF	4CIF60	O	Battery	O

contents. B0 does not support vertical handover, and is not an installed agent. B0 only receives the contents of layer 0, which is streamed as H.264 AVC in the 3G network. B1 fully utilizes the available network bandwidth by moving to the WLAN network and receives a 3.28 average SVC layer. B1 and B1 have an agent, but B1's agent policy is "Network Throughput Preference," and B2's agent policy is "Battery Save Preference." Therefore, B1 receives 0.77 average SVC layer more than B2. But B2 saves 13% more battery power then B1. These results show that by using a streaming agent method, users receive higher quality streaming in a vertical handover environment. Power consumption can also be reduced.

5 Conclusions

This paper proposes a context-aware streaming agent to produce seamless streaming service with optimal quality based on network and terminal capability. Experimental results show the streaming agent improves user satisfaction. By selecting "optimal network conditions", we maximize the 3.28 SVC layer. By selecting "optimal battery condition" policy, we reduce power consumption by 13%. In future work, implementation of a real streaming server, gateway and terminal with the proposed method will be completed along with content quality evaluation, objective measurement of user satisfaction and network performance testing.

References

1. Kuschnig, R., Kofler, I., Ransburg, M., Hellwagner, H.: Design options and comparison of in-network H.264/SVC adaptation. Journal of Visual Communication and Image Representation 19(8), 529–542 (2008)
2. Schierl, T., Wiegand, T., Kampmann, M.: 3GPP Compliant Adaptive Wireless Video Streaming Using H.264/AVC. In: Proc. IEEE International Conference on Image Processing, September 2005, pp. 696–699 (2005)
3. Frojdh, P., Horn, U., Kampmann, M., Nohlgren, A., Westerlund, M.: Adaptive Streaming within the 3GPP Packet-Switched Streaming Service. IEEE Network 20(2), 34–40 (2006)
4. Wu, D., et al.: Streaming Video over the Internet: Approaches and Directions. IEEE Trans. Circuits Sys. Video Tech. 11(3), 282–300 (2001)
5. Horn, U., et al.: Robust Internet Video Transmission Based on Scalable Coding and Unequal Error Protection. Image Commun., Special Issue on Real-time Video over the Internet 15(1), 77–94 (1999)
6. ISO/IEC 21000-7:2004, Information Technology - Multimedia Framework (MPEG-21) - Part 7: Digital Item Adaptation (2004)
7. Reichel, J., Schwarz, H., Wien, M.: Joint Scalable Video Model. JSVM-6. ISO/IEC JTC 1/SC 29/WG 11 (April 2006)

A Message Priority Routing Protocol for Delay Tolerant Networks (DTN) in Disaster Areas

Inwhee Joe* and Sang-Bo Kim

Dept. of Electronics and Computer Engineering, Hanyang University,
Seoul, 133-791 South Korea
iwjoe@hanyang.ac.kr

Abstract. A delay tolerant network (DTN) is a mobile wireless network that is characterized by frequent partitions and potentially long message delivery delays. Compared with conventional networks, the distinguishing feature is that there is no end-to-end connectivity between source and destination. In this paper, we assume that an earthquake has occurred in a city and roads and nodes have been damaged in the disaster. In this situation, we found performance degradation of existing DTN routing protocols due to damage. To address this problem, we propose a DTN message priority routing protocol. In a disaster situation, the proposed protocol was able to deliver more messages than existing DTN protocols, with a lower overhead ratio and lower latency.

Keywords: Routing protocol, message priority, delay tolerant nerworks, disaster area.

1 Introduction

The Internet, first proposed in 1974, has become the global network, made possible in large part due to the TCP/IP protocol and stable infrastructure. In recent years, however, dynamic changes in communication environments and increased user requirements have led to the creation of heterogeneous networks, including wireless communications, satellite or interplanetary communications, and wireless sensor networks. The use of specialized communications techniques makes intercommunication between heterogeneous networks difficult. In order to support communication between heterogeneous networks, the concept of a delay tolerant network (DTN) [1] was proposed. This network structure was created for connections between terrestrial networks with delay times of less than few seconds and outer terrestrial networks with delay times of minutes, hours, days, or even more (e.g., interplanetary networks or satellite communications [2]). However, in recent years, the concept of the DTN was extended to cover so-called opportunistic networks[3], which contain long delay times and no end-to-end path because of frequent changes in connectivity and topology changes due to

* This work was supported by the ITRC Support Program (NIPA-2010-C1090-1021-0009) and the KEIT R&D Support Program of the MKE.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 727–737, 2010.

battery consumption (e.g., sensor networks or vehicular ad-hoc networks). This suggests the suitability of DTNs for environments characterized by malfunction of existing infrastructure, or lack of infrastructure configuration sufficient to disrupt communications. In this paper, we assumed a disaster scenario—specifically, an earthquake in a city—and found that existing DTN protocols were subject to performance degradation in this scenario. To address this problem, we propose DTN message priority routing suitable for emergency situations. This paper is organized as follows. In chapter 2, we introduce stochastic-based routing protocols. In chapter 3, we show the performance degradation of existing protocols in disaster situations and explain how to operate the proposed protocol. In chapter 4, we describe the performance evaluation of the proposed protocol using the ONE simulator. Finally, we conclude in chapter 5.

2 Related Work

DTN routing protocols can be classified as either deterministic or stochastic [4], according to the method of packet delivery. Deterministic routing assumes an environment in which the exact locations and movements of nodes are known. oracle-based [5] and oracular algorithms [6] are typical examples of deterministic routing. Since we consider stochastic environments in this paper, we will not discuss deterministic methods further. Stochastic routing assumes unpredictable movement of each node and no knowledge of the exact location of a node. Messages are forwarded considering the time and place; forwarding decisions are based on mobility patterns, collected data (such as hop count or number of meetings), or additional information. Epidemic [7] and Spray and wait [8] are typical examples of such routing schemes. The long delivery delays in DTN are caused by characteristics of the transmission medium itself, or by lack of stability or mobility of the host. The delay that we mentioned above is also present in mobile networks (e.g. sensor networks, ad-hoc networks, and VANETs). Such existing mobile networks can adopt the concept of DTN. We now introduce some existing DTN routing protocols. Epidemic routing [7] is the simplest and most basic form of DTN routing protocol, and is similar to the flooding algorithm used in wireless ad hoc networks. The Epidemic routing protocol assumes that each node has an infinite storage (buffer) and bandwidth and that every node can store all messages communicated during contact with other nodes. Epidemic routing uses the concept of database replication, in which a node can exchange all of the messages it has in its buffer when meeting another node. Each node also maintains a summary vector to minimize the chance of exchanging duplicate messages. Nodes exchange their summary vector first and then exchange those messages that their respective peers do not have in storage. Epidemic is a practical strategy only in the case of very sparse networks and small messages. Direct delivery routing [13] is an early-generation DTN protocol that was proposed to address storage waste in Epidemic routing. In the Direct delivery, a generated message from the source is not forwarded to any other nodes, but rather is kept until the source meets the destination node. Since it is characterized by message delivery without relay nodes, it has an overhead ratio of 0 and the highest latency time. Prophet [9] is another early-generation DTN protocol. It uses a node's history of previous encounters with neighbor nodes to estimate a probability called delivery predictability. A pair of nodes that have met often in the past have high delivery predictability. When two nodes encounter one another, they exchange vectors containing delivery predictability information which is updated after each contact. Based on this information, each node can select those messages that have high probability of reaching their intended destinations, and transfer them to the contact node. However, this approach does not work well in DTN networks in which node movements are not predictable. Spray and wait [8] is a flooding-based routing protocol which is similar to Epidemic routing. Spray and wait has two protocol states: in the spray phase, a node will flood (spray) a message generated from node itself up to L copy message and, after finishing the flood, shift to the wait phase. In the wait phase, each node maintains its message in the buffer until the destination node is reached. Spray and wait was proposed to reduce excessive packet forwarding (as found in, e.g., Epidemic routing [7]) in networks for the purpose of reducing overhead and buffer overflow. This method is characterized by simple protocol operation, high scalability, and high delivery probability. In Spray and wait, the important parameter is L, the optimum value of which depends on the density, distribution, and mobility profiles of nodes. Analysis to determine appropriate values of L for different network situations is still an open problem. MaxProp [10] protocol is based on the same principles as Prophet [9]. It is designed for delivering messages in a city environment in which network nodes are city buses. The mobility pattern of a bus is a periodic repetition of the same route. MaxProp divides each buffer into two sections. In the first section, stored messages are ordered based on hop count, from low to high. In the second section, messages are ordered by cost, again from low to high. The first section occupies the front of the buffer, and the second section is at the tail. The cost of a path (from source to destination) is calculated using Dijkstra's algorithm.

3 Disaster Situations for DTN

The DTN protocol can be used in many application environments, such as ecological monitoring, interplanetary networks, and communication between heterogeneous networks. Among these, the most suitable environments for DTN are those characterized by malfunction of existing infrastructure, or lack of infrastructure configuration, sufficient to disrupt communications. This is because well organized infrastructure (such as TCP/IP) encompasses various techniques permitting intercommunication with higher performance than DTN affords. In this paper, in accordance with the characteristics of DTN, we considered a disaster situation consisting of an earthquake in a city; we evaluated the performance of existing DTN routing protocols in such a scenario.

3.1 Disaster Scenario

We choose a city for the disaster scenario. In this scenario, approximately 1/3 of roads have been destroyed and 1/3 of nodes (corresponding to cars and pedestrians) have also sustained damage due to the disaster. Of course, the existing infrastructure can no longer perform its basic functions. Specific information is displayed in Table 1. Other conditions in the scenario are the same, with the exceptions of topology damage and number of nodes. For the number of nodes, in order to give a loss of 1/3, we assume damage to each of 20 nodes, or 1/3 of the total, among 60 pedestrians and 60 cars. Among the 20 damaged nodes, we suppose that 10 have lost both transmission ability and movement ability, while the others have lost only their transmission ability.

Table 1. Disaster and Normal Scenarios

Scenario		Normal scenario		Disaster scenario	
Topology damage		none		1/3	
Number of nodes		120 (p60 , c60)		100 (p40, c40, static20)	
Node condition	p	normal	60	normal	40
				unable to move	10
				unable to move unable to transmit (delete from simulation)	10
	c	normal	60	normal	40
				unable to move	10
				unable to move unable to transmit (delete from simulation)	10

(Pedestrians are denoted by P, Cars by C)

3.2 Performance Evaluation of Existing Protocols in Disaster Scenario

To determine the effect on performance of the disaster scenario in Table 1, we configured and simulated the disaster and normal scenarios using the Opportunistic Network Environment Simulator (The ONE) [12]. Figure 1 shows the delivery probabilities for nodes using existing routing protocols, for varying buffer sizes, in the disaster and normal scenarios. The graph for the disaster scenario shows performance degradation with each protocol compared with the normal scenario. MaxProp, which was the best-performing protocol in both scenarios, exhibited a decrease in delivery probability from 89% to 72%. The cause of performance degradation is the loss of roads and nodes due to the earthquake. Since nodes are prevented from choosing the best (shortest) route by road damage, they must choose routes bypassing the disaster area. In this situation, there is no correlation between the destination of a node and that of a message; thus we hypothesize that node loss, rather than damage to roads, is the main cause of poor performance. We now propose DTN message priority routing as a solution to the problem of poor performance in a disaster scenario.

4 DTN Message Priority Routing

In order to achieve better performance in disaster scenarios, we assign priorities to messages. Existing routing protocols, which do not account for disaster situations try to deliver all messages to its destination. In other words, every message has the relationship with horizontal (which means the same message level) not vertical. In harsh environments, such as in disasters, a protocol that attempts to deliver every message will not be as efficient as one that gives greater opportunity to a message with high delivery probability than to a message with low delivery probability.

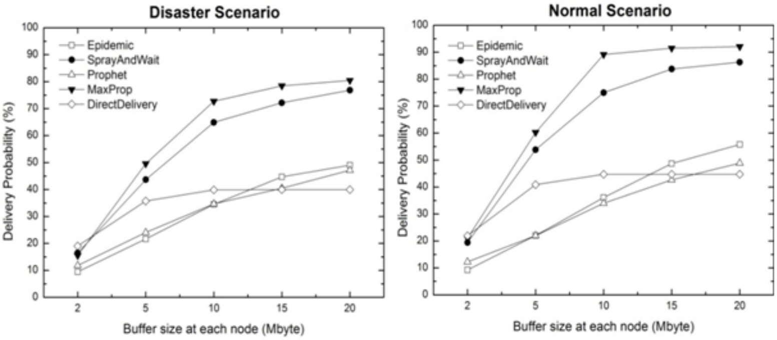


Fig. 1. Total Structure of Spam Filtering System

4.1 Protocol Schematic for DTN Message Priority Routing

DTN message priority routing is a modification of Spray and wait flooding-based routing. The specific protocol operations are as follows.

- Spray phase: remain in this phase until L copies of the message have been disseminated in the network. If L = 6, the number of copies is limited to 6 and forward into the network. When L becomes 0, shift to the Message Priority Forward phase.
- Message Priority Forward phase: In Spray and wait (where the analogous phase is called the wait phase), they would keep the message until the destination node was encountered or until the TTL expired. In the modified protocol, however, we continue with message forwarding according to the type of the message.

- High: Participate in message forwarding if the velocity of the other node is greater than that of this node, i.e., if (where is this node, is the other node, and is the velocity of a node).
- Middle: Do not participate in message forwarding; remain in the wait phase.
- Low: The message is removed from the buffer.

4.2 Selection of Message Priority in Message Priority Forward Phase

In this section, we explain how to select the message type in the Message Priority Forward phase.

1 First, we assume that every node knows its own movement speed, and that every message can be assigned a type of High, Middle, or Low. A newly generated message is assigned Middle as the default type.

2 Every node has a Node Meeting Table (NMT) which records the most recent time at which the node met another node; this table is updated at every contact event. If a node meets the same node again, the NMT is updated with the latest value.

3 We describe an example assuming the following scenario: Node A has a message M13 whose destination is Node C. The NMT of Node B records that Node B met Node C at 7096 seconds. The current time is 8011 seconds.

Table 2. Node Meeting Table of Node B

Node X	Node C	Node G	Node H	.....
1034 s	7096 s	5514 s	1907 s	.....

4 Node A encounters Node B at 8011 seconds. Node A updates its own NMT and refers to Node B's MMT (shown in Table 2) to determine whether or not Node B has met Node C, the destination of M13. Node A learns that Node B met Node C at 7096 seconds. Figure 2 shows the process of referring to Node B's NMT.

5 Using the above facts, we can determine the latest encounter time (LET) of M13 by applying the following equation. This means that Node A will learn how recently Node B met the destination of M13 (i.e., Node C).

$$LET(m) = Time_{current} - Time_{NMT(destination(m))} \tag{1}$$

The LET value is calculated by subtracting the contacted node NMT value associated with the requesting node's message destination from the current time. In our example, the current time is 8011 s, and the NMT value corresponding to the contact time between Node B and Node C is 7096 s. Subtracting, we obtain 915 s (15 minutes). Thus, Node A knows that Node B and Node C (the destination of the message) met only 15 minutes ago. If

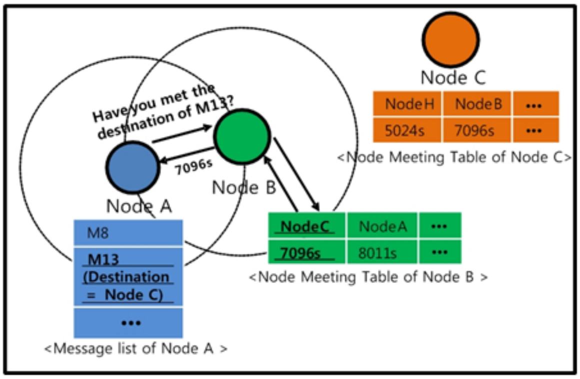

Node C
NodeH
NodeB
•••
5024s
7096s
•••
<Node Meeting Table of Node C>
Have you met the destination of M13?
7096s
Node B
Node A
NodeC
NodeA
•••
7096s
8011s
•••
<Node Meeting Table of Node B >
M8
M13 (Destination = Node C)
•••
<Message list of Node A >

Fig. 2. Node A refers to the NMT of Node B

the value of LET is 60 s (1 minute) or less, we know that it is very close to the destination: the smaller the LET value, in fact, the closer to the destination. The stochastic routing under discussion is available only for prediction of node movement and does not provide exact information about a node's location or time. Therefore, in this paper, in order to forward a message to a neighbor of its destination, we use the LET value. This idea is derived from the FRESH algorithm [11], which tracks the destination node using the most recent encounter time in the absence of location information, such as GPS.

6 After the LET value has been determined, a message will be assigned a type of High, Middle, or Low, depending on the threshold value, as shown in Figure 3. The amount of message forwarding (High) is dependent on the size of the threshold value. The "Low" type will be assigned when the LET value is greater than TTL (time to live) and the buffer utilization of the node exceeds 80%.

7 Each node repeats the above steps upon coming into contact with another node; the message type of each message held by a node will be assigned dynamically.

If the type of a message is High, then the node compares its own velocity with the contacted node's velocity and, if the contacted node is faster, participates in message forwarding. When the velocity of a node is increasing, the node meets other nodes more often. Therefore, we consider the velocity of a node for forwarding, (packets).

$$v = \frac{s}{t} \tag{2}$$

In other words, if , according to (2), the shifting distance of b is longer than that of a (where is velocity, is time, and is shifting distance). By considering velocity, we can reduce the overhead ratio and also minimize degradation of the delivery probability.

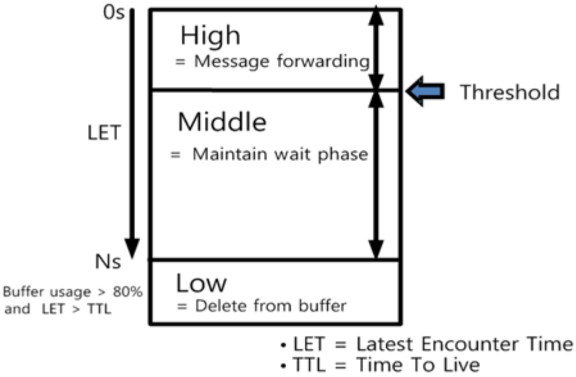


Fig. 3. Selection of message priority

4.3 Selection of Threshold Value

The optimum threshold value for each buffer size for a node is determined experimentally; these values vary depending on node conditions and topology. The optimum threshold value is defined to be that providing the best delivery probability with low overhead ratio. The optimum value is shown in Figure 4. As the size of the buffer increases, the optimum value also increases; each threshold value in the disaster scenario is greater than those in the normal scenario.

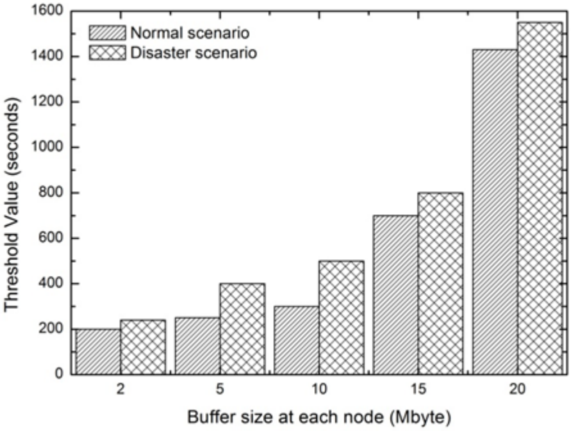


Fig. 4. Selection of the optimum threshold value from the simulation

4.4 Improved Buffer Utilization Using Acked Message

When a message arrives at its destination, any copies of the message forwarded into the network become useless. These copies are only terminated upon TTL expiration; this is a cause of inefficient buffer utilization. In this paper, in order to remove copies of delivered messages before TTL expiration, we use Acked Message [10] at every node contact. Thus, we can remove copies of delivered message sooner than by TTL expiration, maximizing buffer utilization.

5 Performance Evaluation

We evaluated the performance of the proposed routing protocol using The Opportunistic Network Environment Simulator (The ONE) [12], developed by the University of Helsinki, Finland. The simulation environment is shown in Table 3. In order to evaluate the proposed protocol, we measured three elements for each of two scenarios. The first element is the delivery probability, that is, the success rate for deliveries. The second is the overhead ratio, calculated as the rate of undelivered messages per completed delivery. The final element is the average latency. For mobility models, we adopted shortest-path map-based movement, which is designed to move to the destination along shortest path. The characteristics of mobility models ensure that a node will find the next shortest path as an alternative in case of shortest path failure in a disaster scenario. We compared the results with MaxProp, which showed the best performance in the disaster scenario.

Table 3. Simulation Environment

Scenario	Disaster , Normal
Map size (m)	width: 4500, height: 3400
Number of nodes	100 (p40, c40, static20), 120 (p60, c60)
Simulation time (h)	12 (43200 s)
Mobility model	Shortest-path map-based movement
Node speed (m/s)	P (0.5–1.5), C (2.7–13.9), Static (0)
Transmission range (m)	10
Transmission speed (bps)	250k
Buffer size (byte)	2M, 5M, 10M, 15M, 20M

Figure 5, parts (a), (b), and (c) show the results in the disaster scenario, and (d), (e), and (f) the results in the normal scenario. The proposed protocol is denoted “SnMPF”. Comparing part (a) with part (d), the proposed protocol shows similar delivery probability to that of MaxProp in the normal scenario. However, in the disaster scenario, the proposed protocol shows higher performance than that of MaxProp. In other words, the proposed protocol is suitable for disaster situations. It can be seen in Figure 5 (b) and (e) that the overhead ratio of the proposed protocol is similar with the Spray and wait. Moreover, the overhead

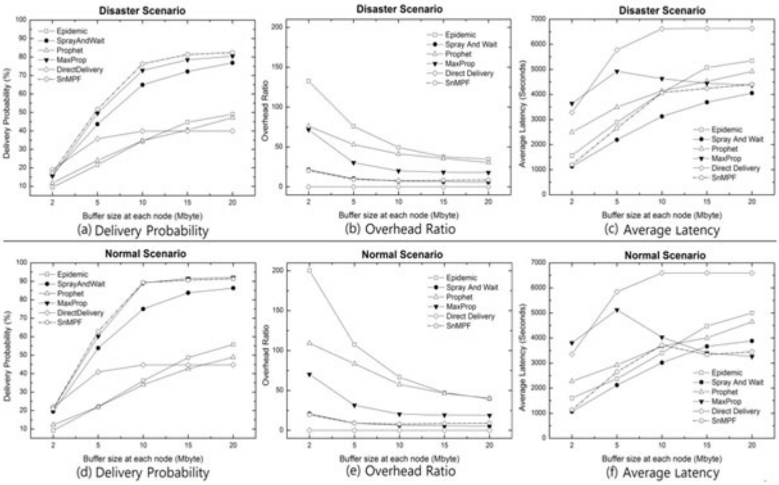

(a) Delivery Probability
(b) Overhead Ratio
(c) Average Latency
(d) Delivery Probability
(e) Overhead Ratio
(f) Average Latency

Fig. 5. Simulation results in disaster and normal scenarios

ratio of MaxProp with a 2 Mbyte buffer size is 71.6, as compared to 20.5 for the proposed protocol, i.e., approximately 3.4 times lower for the proposed protocol than for MaxProp. Finally, Figure 5 (c) and (f) show increased overall average latency in the disaster scenario. The average latency for the proposed protocol is 1196 seconds with a 2 Mbyte buffer size in the disaster scenario; MaxProp, however, has a latency of 3641 seconds. Thus, latency for the proposed protocol is approximately 3 times lower than for MaxProp.

6 Conclusions

In this paper, we describe performance degradation of existing protocols in disaster situations. In order to resolve this problem, we propose a DTN message priority routing protocol, which is a modification of the Spray and wait flooding-based routing protocol. The proposed protocol is designed to give greater opportunity to a message that has high delivery probability than to one with low probability, for efficient routing in disaster situations. The results were evaluated using the ONE simulator. The outcome was higher delivery probabilities in the disaster scenario, along with lower overhead ratios. Finally, the average latency is also lower than that of MaxProp in the disaster scenario.

References

1. Delay tolerant networking research group, http://www.dtnrg.org
2. Burleigh, S., Hooke, A., Torgerson, L., Fall, K., Cerf, V., Durst, B., Scott, K.: Delay-tolerant networking:an approach to interplanetary internet. IEEE Communications Magazine 41, 128–136 (2003)

3. Pelusi, L., Passarella, A., Conti, M.: Opportunistic Networking: data forwarding in disconnected mobile ad hoc networks. IEEE Communications Magazine issue on Ad hoc and Sensor Networks 44(11) (November 2006)
4. Zhang, Z.: Routing in Intermittently Connected Mobile Ad Hoc Networks and Delay Tolerant Networks: Overview and Challenges. IEEE Communications Surveys & Tutorials (2006)
5. Jain, S., Fall, K., Patra, R.: Routing in a Delay Tolerant Network. In: SIGCOMM 2004 (2004)
6. Handorean, R., et al.: Accommodating Transient Connectivity in Ad Hoc and Mobile Settings. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 305–322. Springer, Heidelberg (2004)
7. Mitchener, W., Vadhat, A.: Epidemic Routing for Partially Connected Ad hoc Networks. Technical Report CS-2000-06, Duke Univ. (2000)
8. Spyropoulos, T., Psounis, K., Raghavendra, C.S.: Spray and wait: An Efficient Routing Scheme for Intermittently Connected Mobile Networks. In: Proc. ACM WDTN, pp. 252–259 (August 2005)
9. Lindgren, A., Doria, A., Schelen, O.: Probabilistic Routing in Intermittently Connected Networks. In: Proc. SAPIR Workshop, pp. 239–254 (August 2004)
10. Burgess, J., Gallagher, B., Jensen, D., Levine, B.N.: MaxProp: Routing for Vehicle-Based Disruption- Tolerant Networks. In: Proc. IEEE Infocom (April 2006)
11. Dubois-Ferriere, H., Grossglauser, M., Vetterli, M.: Age matters: dfficient route discovery in mobile ad hoc networks using encounter ages. In: Proceedings of ACM MobiHoc (2003)
12. The ONE simulator, http://www.netlab.tkk.fi/tutkimus/dtn/theone/
13. Spyropoulos, T., Psounis, K., Raghavendra, C.S.: Single-copy routing in intermittently connected mobile networks. In: Proc. Sensor and Ad Hoc Communications and Networks SECON, pp. 235–244 (October 2004)

A Queue Management Algorithm for QoS Provisioning in WMSN

Junghye Kim[1], Seongwon Myung[1], Yongjae Jang[2], and Sungkeun Lee[2,*]

[1] Dept. of Computer Eng., Sunchon national University, Korea
jlove79kim@hanmail.net, floydwon@gmail.com
[2] Dept. of Multimedia Eng., Sunchon national University, Korea
elsv1114@gmail.com, sklee@sunchon.ac.kr

Abstract. Wireless Multimedia Sensor Network (WMSN) is an advanced concept of general wireless sensor network with multimedia sensing function. The development of WMSN is getting realized in accordance with the increased demands to provide multimedia data transmission services based on wireless sensor networks. It requires a traffic control mechanism that can support Quality of Service(QoS) to handle different services efficiently. This paper classifies traffic in WMSN as periodic monitoring traffic, event traffic, multimedia traffic and query-based traffic. This paper proposes traffic control mechanism that guarantees differentiated QoS in regard to type and feature of WMSN traffic and conducts performance analysis for our mechanism.

1 Introduction

Wireless Sensor Network (WSN) is composed of networks that comprise of sensors which monitor static physical phenomena such as temperature, pressure, humidity and location and transmit them to sink nodes [1].

With rapid development of computer and wireless network technology, WSN is in the center of topics for many fields such as national defense, fire disaster monitoring, environmental exploration, surveillance system, traffic system, health care, monitoring service system, and manufacture automation.

In addition, as recent development for low cost hardware such as CMOS camera and microphone become flourished, Wireless Multimedia Sensor Network (WMSN) which enables sensing video and audio data, still images and static sensor data is required to be deployed immediately.

While WSN only demands low calculation capability, WMSN which catches, restores and transfers multimedia data requires capabilities for fast calculation and for massive data transmission. While related studies of WSN consider all data as the same type, WMSN takes traffic as four types with distinctive feature in each. Each type of traffic includes its loss, latency and tolerance so allowing similar loss and latency rate to all traffics is not appropriate in terms of resource utilization; entire performance would decline eventually.

[*] Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 738–750, 2010.

This study proposes traffic control mechanism that guarantees differentiated QoS in regard to type and feature of WMSN traffic and conducts performance analysis for our mechanism.

Section 2 in this paper shows analysis of existing studies related to the topic and section 3 describes proposed protocol in detail. Section 4 reports proposed protocol performance and result analysis. Last, we will bring our conclusion in section 5.

2 Related Work

Wireless Multimedia Sensor Network is an advanced concept of general wireless sensor network with multimedia sensing function. Recent advancement of low cost hardware such as CMOS camera and microphone promotes to realize Wireless Multimedia Sensor Network (WMSN) that enables sensing video/audio data, still images and static sensor data. While researches on protocol development in WSN heavily concentrate on energy efficiency, studies in WMSN not only consider energy efficiency but also emphasize the importance on QoS guarantee. Recent researches on QoS guarantee in WSN are introduced as follow.

SPEED [4] establishes path by calculating speed of edge-to-edge transmission latency and maximum transfer speed, skipping the edge-to-edge path setup process. When congestion occurs, it guarantees service quality using back-pressure and re-routing technologies. Sequential Assignment Routing (SAR) [5], the first research for service quality in WSN, is an approach to reflect fixed priority to a network. As SAR utilizes table driven multi-path method with fixed priority until it arrives to sink node, it determines its path dynamically with regard to requirement of latency and bandwidth control.

Multi-path Multi-Speed Routing Protocol (MMSPEED) [6] is a protocol that transfers packets by assigning differentiated priority. It guarantees reliability through Multi-path and provides timeliness with multiple network-wide packet transmission speed.

Along with those protocols, other protocols has been introduced such as an Enery-aware QoS routing protocol [7] and RAP (A Real-time communication Architecture for large-scale wireless sensor networks) [8].A general mechanism that provides reliability in WSN is multi-path routing. Transmitting copies of the same data thru multiple paths, it is a concept that at least more than one copy can be delivered to sink node in time.

3 Proposal of Protocol

3.1 Traffic Control Framework

Most data pattern in WMSN is a many-to-one type that multiple sink nodes deliver data to a single sink node. Each sensor node not only works as a source that generates sensing data but also performs as a router that broadcasts received data to sink node. Traffic control mechanism for QoS guarantee in WMSN consists of packet marking algorithm and queue management mechanism. Traffic control mechanism in WMSN that we consider in this paper is illustrated in Fig. 1.

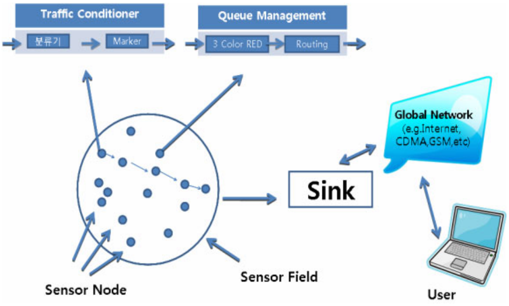


Fig. 1. Traffic control function of sensor nodes

Packet marking algorithm judges a packet with its importance, marks one of Green, Yellow and Red colors with the importance then transfers to sink node. Queue management mechanism, a function of broadcasting node, executes determination for whether a packet is delivered or dropped based on transfer queue's occupancy rate. Engaging both marking algorithm and queue management mechanism, it can guarantee differentiated service quality in WMSN according to each traffic pattern.

3.2 Traffic Categorization

Most existing protocols consider all traffic types as having the same feature. Meanwhile, in WMSN traffic can be categorized as periodic monitoring traffic, event traffic, multimedia traffic, query-based traffic in accordance with features of application service. In this paper, when transferring data it transmits data by designating service quality pattern and service quality level based on traffic pattern. Service quality pattern is categorized as energy efficiency, latency sensitive, reliability and transfer rate and service quality level as Green, Yellow and Red.

Service quality pattern and service quality level responding to traffic pattern are designated by the standard below.

Periodic monitoring traffic is a traffic that monitors circumstantial information such as temperature or humidity in certain period then sends them to sink node. When current data does not differ significantly to the previous data, it designates its service quality level considering energy efficiency rather than conceiving latency or reliability. When current data shows significant difference to the previous data, it considers latency or reliability to designate its service quality level.

Event driven traffic, that is to respond emergency event such as a forest fire detection or intrusion detection immediately, is a type of traffic that is transferred as the highest service quality level so that it is delivered instantaneously to sink node with the highest reliability and the minimal latency.

Multimedia traffic is a traffic that continuously delivers video data from CCTV or audio data from sound capture hardware. Key frame data in this traffic is designated as higher service quality level to focalize its latency and reliability and other data such as supplementary frame or differential frame is delivered as relatively lower service quality level.

Query-based traffic is the response traffic from a sensor node that is requested to respond to a sink node when a sink node generates query to the sensor node. When a sink node sends query, the sink node assigns reliability level to determine service quality level of response data which will come to the sink node.

In this paper, it utilizes Type Field to distinguish traffic type and Priority Field to represent each packet's priority. Green packet, the highest priority, is guaranteed to send first and is supposed to be delivered in time at all costs. Red packet, the lowest priority, is dropped first when congestion occurs. Red packet, even though lost in congestion, includes data which can be delivered thru other packets which sink node receives. Yellow packet gets middle-class service in between Green and Red packet.

3.3 Marking Algorithm in Source Node

3.3.1 Periodic Monitoring Traffic

Periodic monitoring traffic has a feature that when a particular environment does not show any significant change, the value of data does not often change. Yet, even when the value of data does not change, it requests data transfer. This traffic type is good when guaranteeing reliability and latency at certain point. Also, when current data, which is barely different to previous data, will be lost or experience delay during congestion in a network, it does not affect to entire network reliability. Therefore, if current data is not significantly different to previous data, energy efficiency rather than latency or reliability will be the factor for determining the packet’s service quality level. If current data is substantially different to previous data, designating service quality level to guarantee latency and reliability and sending the data will secure energy efficiency without allowing serious effect to application service.

When data to transfer occurs, source node compares current packet to both one cyclic previous packet and two cyclic previous packet then marks Priority Field as Red when the difference among them is smaller than α. If current packet does not show significant difference to one cyclic previous packet but the difference to two cyclic previous packet is bigger than α, it means that there is a difference between current packet and two cyclic previous packet and Priority Field will be marked as Yellow. If current packet differs to one cyclic previous packet more than α, it marks Green then sends the packet.

Periodic monitoring packet is a type of traffic that monitors environmental data such as temperature or humidity at seconds or tens seconds units and transfers them to sink node. This traffic does not change often when certain environmental data do not change. However even when there is no change of certain environmental data, it demands transferring. Although this type of traffic is optimal if reliability and delay can be guaranteed at certain level, when network congestion occurs it does not harm the entire network reliability if a data, which is not much different to the previous data, is lost or delayed. Simply when a data, that is much different to the previous data, is lost, it causes delay on seizing a sign or foreshadow of emergency event occurrence, resulting serious problems.

Hence, when a data is not much different, its QoS level should be setup based on energy efficiency rather than delay or reliability. When a data is much different, its QoS level should be designated to guarantee delay and reliability so that it does not affect serious effect to application service and secures energy efficiency. Depending on the difference between the previous data and newly estimated data(current data), packet priority can be marked. Our proposed marking algorithm is shown in Fig 3. Source node establishes its path by Priority Field. If packets are only considered to be transferred in the shortest path, network congestion occurs as too many packets are driven to the shortest path, resulting many resources of nodes located out of the shortest path on hold. Therefore, resource utilization becomes inefficient because one side of the entire network experiences congestion, which ensues packing dropping consistently, harming reliability and timeliness and the other side of the network becomes waiting states, remaining their resources unused.

Our proposed protocol delivers 'Green' packets at shortest path and 'Yellow' and 'Red' packets at paths other than the shortest. In this case, we might think that it would not fit to timeliness since 'Yellow' and 'Red' packets are not transferred via shortest path. The result demonstrated a better timeliness as these two type of packets utilize relatively less occupied paths which occur less Queueing delay and processing delay on the middle nodes. As the middle nodes are idle, reliability of entire networks becomes also very efficient. In this way, when 'Yellow' and 'Red' packets utilize extra paths, resources of nodes in the shortest path for 'Green' packet can take advantage in terms of the entire network performance.

3.3.2 Event Traffic

Event traffic is a traffic that is designated as the highest service quality level to provide the utmost reliability and the minimum latency. As one of features in WMSN is high node density, multiple nodes, when an event occurs, catch similar values then try to transfer data to sink node. Thus, higher priority packets burst into network as data duplication also occurs and each sensor node sends packets with higher priority.

Eventually, part of or entire network will experience congestion. Although, this data duplication might be able to reduce the amount of data that is delivered thru data synchronization process in the middle nodes, the load to data synchronization process will be expected costly.

If it recognizes that multiple sensor nodes duplicatedly detects an event occurred, only a few nodes of them will transfer data and the rest will not transfer. If all the nodes that detect an event begin to transfer with not marking all the packets as higher priority, it can reduce the amount of traffic infused into network and immediate treatment can be done against congestion state as it cuts down packets with higher priority.

The proposed protocol in this paper checks how many nodes have been detected an event occurred and starts to mark packets as Green, Yellow and Red discriminately.

E_num presents the number of node that senses the same event concurrently and Pro is a variable that is to mark packet's Priority Field with probability by generating a random number. We assume that the number of nodes that senses the same event occurred concurrently is acknowledged.

Header nodes will be marked as Green unconditionally. Nodes other than header node will get their Priority Field marked in proportion to different ratio from the number of nodes that sense the event.

```
data(t-2T) : Data estimated at 2 cycle-periodic unit
data(t-T) : Data estimated at previous periodic unit
data(T) : Data current estimated
Red_count : The number of packets that are marked as RED continuously.
Red_THRE : maximum number of packet that are marked as RED continuously

Initialize :
Red_count = 0;
Red_THRE = $\alpha$;
    When Period Monitoring Data are generated :
       if(|data(t-T) - data(T)| / max(data(t-T),data(T))<$\alpha$) {
         if (| data(t-2T) - data(T)| / max(data(t-2T),data(T))< $\alpha$) {
            Priority_Field = "RED";
            Red_count++;
         }
         else {
            Priority_Field = "YELLOW";
            Red_count = 0;
         }
       }
       else {
            Priority_Field = "GREEN";
            Red_count = 0;
       }
   }
```

Fig. 2. Algorithm for periodic monitoring traffic

```
E_num : Number of nodes that sense the same
event
Pro : a variable gained from random number

If (I am Header node)
    Event Data Priority Field = "Green";
else if (E_num > 6) {
    Pro = random();
    if (Pro<= ) Priority Field = "Green";
    else if ( <Pro<= ) then
        Priority Field = "Yellow";
    else Priority Field = "Red"; }
else if (2<E_num<=6){
    Pro = random;
    if (Pro<= ) Priority Field = "Green";
    else if ( <Pro<= ) then
        Priority Field = "Yellow";
    else Priority Field = "Red"; }
else Priority Field = "Green";
```

Fig. 3. Event traffic marking algorithm

If the number of nodes is more than 3 and less than 6, it raises Green ratio more than that of which the number of nodes is more than 6 because if marking Priority Field as Green as the same ratio of that of more than 6 nodes, only one or two packets including header node will be marked as Green. If this Green packet will not be lost until sink node receives in harsh network condition, it will not be worried. However it is very dangerous to send only one Green packet since Green packet might be lost if harsh congestion occurs in the middle nodes. In case if the amount of nodes that sense the event is less than 2, it unconditionally marks all packets as Green.

3.3.3 Multimedia Traffic

Multimedia Traffic represents video, image and audio data. With the nature of multimedia, the difference between the front frame and the rear frame is small. Hence, if the entire data of each frame will be transferred, packets with the same content will be cloned and add the network load as the cloned packets will attempt to go forward. Also, even when one frame in a multimedia data is not sent 100%, there is no problem for a user to verify the content delivered.

As a result, multimedia traffic can reduce network load hence guarantee higher reliability and timeliness by assigning differentiated priority to each frame to send them. I frame is marked as Green to guarantee loss-free service and P frame and B frame are marked as Green, Yellow and Red proportionally. In this paper, we mark 20% of packets as Green, 50% as Yellow and 30% as Red.

3.3.4 Query-Based Traffic

Query-based traffic, which is very different from other traffics introduced, is the response when sink node requests to certain sensor node.

The weakness of WMSN is limited energy and slow calculation capability of sensor node. Its strength, on the other hand, is unlimited energy and fast calculation capability of sink node. Using this feature, when sink node demands query, the sink node attaches priority of response packet for the query. Sensor node, when responding to the query, marks priority of response packet as the same as priority designated from sink node without additional calculation process.

3.4 Establishing Initial Path at Source Node

Source node sets path based on marked priority field. If packets are sent only via shortest path like existing protocols, packets are prone to gather into nodes inside the shortest path, generating network congestion. Nodes other than nodes inside the shortest path will be on hold, making inefficient resource utilization.

In this paper, Green will be sent via shortest path and Yellow and Red packets are transferred thru other paths. In this case, it might think that it is not timely since Yellow and Red packets do not use shortest path. However, as far as the packets are not gathering to certain path, it will bring a better performance in terms of timeliness eventually because the packets will utilize relatively idle path, resulting in lowering

Queuing Delay and Processing Delay at the middle nodes. Also, because the middle nodes are idle, it is very efficient in reliability perspective as the number of lost packets will be diminished. In the end, the entire transfer rate will increase.

3.5 Queue Management Mechanism at the Middle Node

All nodes employ only one queue. In the middle nodes, packets should be dropped or transferred in proposition to congestion level using RIO algorithm [10]. When a packet comes, it checks its congestion level then determines whether the packet will be dropped. Congestion level can be checked by queue occupancy ratio. Congestion level is categorized as three levels below.

- normal operation
- congestion avoidance
- congestion control

Normal operation, a queue state lower than threshold, is that there is no congestion and packet transferring is placid. In this level, all the packets are transferred without loss. Congestion avoidance is a state of pre-condition to reach congestion by the number of transferring packet which increases queue occupancy ratio higher than threshold. This drops packets proportionally and does transfer only a few incoming packets. Congestion control is a state of queue occupancy ratio as 1, fully occupied. Judging if congestion occurs, it drops all the incoming packets.

Once it decides to transfer packets, the next step will be the selection of next node toward sink node. When queue occupancy ratio is less than α, all the packets will be delivered via optimal path and when queue occupancy ratio is higher than α, Green packets are delivered thru optimal path and Yellow and Red packets are sent via either optimal path or other paths by generating random function. Fig. 4 demonstrates operation of queue management mechanism.

avg_queue in Fig. 4 algorithm presents current queue occupancy rate. P_drop is a variable to determine whether incoming packet will be dropped or transferred proportionally. P_routing is a variable to select the shortest path and spare paths.

When P_(color) is smaller than min_(color), 100% of Green packets are delivered via the shortest path and 70% of Yellow packets are sent thru the shortest path and 30% of the remaining packets sent thru spare paths. If P_(color) is in between min_(color) and max_(color), it drops packets proportionally. Undropped packets in this case, 70% of Green packets will be sent via the shortest path and 30% of them will be delivered thru spare paths. All Red packets are sent thru spare paths.

In any case, Green has a priority to be transferred via the shortest path. However, as the shortest path is assigned not only Green packet but also Yellow packet, it guarantees Yellow packet throughput at a steady rate. Utilizing both the shortest path and spare paths, it can ensure service quality with timeliness, reliability and energy efficiency.

```
avg_queue : Current Queue occupancy rate
P_drop : a variable to drop packets randomly with probability P
P_routing : a probability variable that determines shortest path
and spare paths

if (Priority Field = "Green"){
    avg_green = calculate Green queue size;
    calculate probability P_green;
    if (P_green < min_green)
        forwards packet to shortest path;
    else (min_green < P_green<max_green)
    {   P = random();
        if (P_green<P)
                forwards packet to shortest path;
        else drop this packet; }
    else drop this packet; }
else if (Priority Field = " Yellow" )
{   avg_yellow = calculate Green queue size & Yellow queue size;
    calculate probability P_yellow;
    if (P_yellow < min_yellow) forwards packet;
    else (min_yellow < P_yellow<max_yellow)
    {   P = random();
        if (P_yellow<P) forwards packet to path;
        else drop this packet; }
    else drop this packet; }
else
{   avg_red = calculate average queue size;
    calculate probability P_red;
    if (P_red < min_red) forwards packet to path;
    else if (min_red < P_red<max_red)
{   P = random();
    if (P_red<P) forwards packet to path;
    else drop this packet; }
else drop this packet; }
```

Fig. 4. Queue management mechanism

4 Performance Analysis and Result

4.1 Experiment Configuration

Performance analysis for our proposed mechanism has conducted using simulation method. Fig 5 depicts node structure design in our simulation.

Our simulation carried on under Intel Core 2 Quad Q9300 2.5Hz, a CPU, RAM 2GM, Fedora Core 10, Kernel 2.6.27.38, gcc 4.3.2-7, an operating system.

Service time is average 100ms and packet generation is setup as exponential distribution. Periodic traffic occurs every one second and event driven traffic comes out average five seconds with following Poisson distribution by generating random function. Multimedia traffic carries a low resolution MPEG2 video data at 100kbps per second. Base traffic, FTP traffic, sends at 100kbps per second. Total nodes are nine; 4th node is header node.

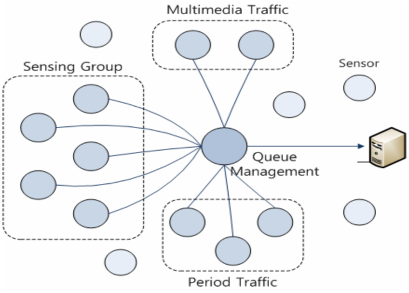


Fig. 5. Configuration of nodes

Although actual field generates periodic traffic, event driven traffic and multimedia traffic from all nodes, to get a more accurate analysis of result, we, in this paper, divide them to endow each role. We assume that from 1st to 3rd nodes produce periodic traffic, from 4th to 8th nodes generates event driven traffic and 9th node transfers multimedia traffic. Our simulation compares both No Marking Droptail FiFo Queue mechanism, an existing mechanism, and 3 Color Marking 3 Color RED Queue mechanism, proposed in this paper. When each node in the simulation generates each traffic type, we conducted performance analysis by comparing arrival rate on each priority of traffic.

4.2 Simulation Result and Analysis

Comparing total transfer rate to each mechanism failed to show significant difference; 50.31% on No Marking Droptail FiFo Queue Mechanism and 47.08% on 3 Color Marking 3 Color Queue Mechanism. We demonstrated the comparison of both mechanisms in terms of priority of packet in Fig 6. First, as No Marking Droptail FiFo queue mechanism do not establish priority, transfer rate of Green packet is 41.98%, transfer rate of Yellow packet as 40.32% and transfer rate of Red as 57.75% with total average rate from 40 ~ 50%. Even Red packets whose total amount outnumbers the rests showed the highest transfer rate. 3 Color Marking 3 Color RED queue mechanism, though it might show lower performance than the previous one on total transfer rate, demonstrated that transfer rate of Green packet is 98.18%, transfer rate of Yellow packet is 30.08% and transfer rate of Red packet is 29.9% in terms of priority, showing that 47.08% of total transfer rate is allocated to Green packet.

Comparing transfer rate of each node, as is the same as total transfer rate, our proposed mechanism guarantees high transfer rate on Green packet. Yet, existing mechanism still showed close results among Green, Yellow and Red on the whole.

As we compare each traffic, the result demonstrated that header node of event driven traffic is 32% guaranteed in existing mechanism while it is 99% guaranteed in proposed mechanism.

Also, comparing the two mechanisms whether all the triggered events are sent to sink node when an event triggers, though the existing mechanism showed similar transfer rate in general, it dropped nine out of 100 event sequences in our simulation.

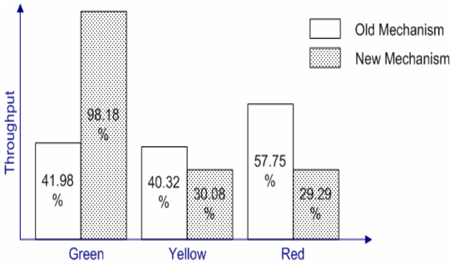


Fig. 6. Comparison of throughputs according to packet priority

Even though five nodes caught sensing the event and sent them, a few sequences were duplicated up to five times then sent to sink node. While other few sequences could not be transferred and, consequently, dropped,, it is because the middle nodes drops regardless of priority of packet. This can cause serious problems to the network with feature of event traffic. Yet, in the proposed mechanism, though total transfer rate of event traffic is the same as that of existing mechanism, all the triggered event were broadcasted to sink and an event that was dropped in header node were sent to other node since header nodes were 99% delivered to sink node. Finally, about 100 triggered events were all transferred to sink and frequency of duplicate transfer for all sequences was held three times for a single event. Strengths of proposed mechanism in event traffic are that, first, all the triggered events were sent to sink node and, second, at least three duplicated data for one event are delivered. Even when multiple nodes simultaneously sense an event, the sensed data can be infinitesimally different according to the location of each node. This difference can be useful information to explore the location of event occurrence and its situation. Therefore, with second strength described above, it can provide more specific information about the triggered event to a user. In case of multimedia traffic, existing mechanism provide 97% on Green packet, 24% on Yellow packet and 15% on Red packet. Our proposed mechanism in this paper marks key frames as Green, showing almost 100% transfer rate of these key frames. The transfer rate of supplement frames, B and P frames, is 24% and 15% each which are relatively low. While measurement of the simulation result did not show any difference between both mechanisms on total transfer rate, our proposed mechanism succeeded to deliver Green packets more hence provide a better service quality to users.

5 Conclusions

Wireless Sensor Network with capability of multimedia data consists of data sensor node which perceives sound/motion and video sensor node that captures interesting event video data. With the maturity of hardware and power supply technology, this WMSN will be constructed and will be applied to many fields. Unlike other sensing data, multimedia data necessitates massive quantity of calculation and transmission; existing sensor networks cannot accept multimedia data. To utilize limited resource in sensor network, this paper analyzes traffic type in WMSN and proposes marking algorithm and queue management mechanism to provide differentiated service quality in relation to latency, energy efficiency and reliability on each type of traffic.

From the result of simulation analysis, total transfer rates on existing mechanism and our proposed mechanism are 50.31% and 47.08% each, showing no significant difference. However, existing mechanism provides close transfer rate to Green packet (41.98%), Yellow packet (40.32%) and Red packet (57.75%). This means that the limited resource is allocated even regardless of significance (priority) of each packet. Therefore, when an important packet (high priority packet) is dropped, relatively less high priority packet can gain some of limited network resource.

On the other hand, our proposed mechanism, allowing 98.1% of Green packet, 30.08% of Yellow and 29.29% of Red packet, guarantees high service quality to high priority packet. If network resource is enough to accept all the incoming data, the proposed mechanism do not show significant different to existing mechanism by allowing all the incoming packets to share network resources. Yet, if network starts to experience congestion, existing mechanism evenly provides network source to any packet while the proposed mechanism saves high priority packets to transfer and drops low priority packets. Eventually, even when transferring the same amount of data, it gives enhancement of service quality on entire network resource utilization by guaranteeing service quality to important packets.

The proposed mechanism in this paper assumes that it has the number of nodes that sense an event from Event driven traffic at the same time. Therefore, further research for figuring out the number of nodes that sense an event simultaneously should be done for the support of this paper.

Acknowledgement

“This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)” (NIPA-2010-(C1090-1021-0009))”.

References

Akylidiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: A Survey on Sensor Networks. IEEE Communication, Magazine (2002)

Akyildiz, I., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: A survey. Computer Networks 38(4), 393–422 (2002)

Akyildiz, I.F., Melodia, T., Chowdury, K.R.: Wireless Multimedia Sensor Networks: A Survey. IEEE Wireless Communication 7(32-39), 1536–1284

He, T., Stankovic, J., Lu, C., Abdelzaher, T.: SPEED: A stateless protocol for real-time communication in sensor networks. In: Proc. 23rd International Conference on Distributed Computing Systems, May 2003, pp. 46–55 (2003)

Sohrabi, K., Gao, J., Allawadhi, B., Pottie, G.: Protocols for self-organization of a wireless sensor network. IEEE Pers.l Commun. 7(5), 16–27 (2000)

Felemban, E., Lee, C.-G., Ekici, E., Boder, R., Vural, S.: Probabilistic QoS guarantee in reliability and timeliness domains in wireless sensor networks. In: Proc. IEEE INFOCOM, March 2005, pp. 2646–2657 (2005)

Akkaya, K., Younis, M.: An energy-aware QoS routing protocol for wireless sensor network. In: Proc. Workshops in the 23rd International Conference on Distributed Computing Systems, May 2003, pp. 710–715 (2003)

Lu, C., Blum, B., Abdelzaher, T., Stankovic, J., Tian, H.: RAP: A real-time communication architecture for large-scale wireless sensor networks. In: Proc. IEEE Real-time Systems Symposium(RTSS), December 2001, pp. 55–66 (2001)

Deb, B., Bhatnagar, S., Nath, B.: ReInForM: Reliable information forwarding using multiple paths in sensor networks. In: Proc. 28th Annual IEEE International Conference on Local Computer Networks, Bonn, Germany, October 2003, pp. 406–415 (2003)

Network Simulator (NS), University of California at Berkeley, CA (1997), `http://isi.edu/nsnam/ns`

OLAP Data Cube Compression Techniques: A Ten-Year-Long History

Alfredo Cuzzocrea

ICAR-CNR and University of Calabria
I-87036 Rende, Cosenza, Italy
cuzzocrea@si.deis.unical.it

Abstract. *OnLine Analytical Processing* (OLAP) is relevant for a plethora of *Intelligent Data Analysis and Mining Applications and Systems*, as it offers powerful tools for exploring, querying and mining massive amounts of data on the basis of fortunate and well-consolidated multidimensional and a multi-resolution metaphors over data. Applicative settings for which OLAP plays a critical role are manyfold, and span from *Business Intelligence* to *Complex Information Retrieval* and *Sensor and Stream Data Analysis*. Recently, the Database and Data Warehousing research community has experienced an explosion of OLAP-related methodologies and techniques aimed at improving the capabilities and the opportunities of complex mining processes over heterogeneous-in-nature, inter-related and massive data repositories. Despite this, open problems still arise, among which the so-called *curse of dimensionality problem* plays a major role. This problem refers to well-understood limitations of state-of-the-art OLAP data processing techniques in elaborating, querying and mining multidimensional data when data cubes grow in size and dimension number. This evidence has originated a large spectrum of research efforts in the context of *Approximate OLAP Query Answering* techniques, whose main idea consists in *compressing target data cubes in order to originate compressed data structures able of retrieving approximate answers to OLAP queries at a tolerable query error*. This research proposes an excerpt of a ten-year-long history of OLAP data cube compression techniques, by particularly focusing on three major results, namely $\Delta - Syn$, K_{LSA} and $\mathcal{LCS} - Hist$.

1 Introduction

OnLine Analytical Processing (OLAP) [6] is relevant for a plethora of *Intelligent Data Analysis and Mining Applications and Systems*, as it offers powerful tools for exploring, querying and mining massive amounts of data on the basis of fortunate and well-consolidated multidimensional and a multi-resolution metaphors over data. Applicative settings for which OLAP plays a critical role are manyfold, and span from *Business Intelligence* to *Complex Information Retrieval* and *Sensor and Stream Data Analysis*. Recently, the Database and Data Warehousing research community has experienced an explosion of OLAP-related

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 751–754, 2010.

methodologies and techniques aimed at improving the capabilities and the opportunities of complex mining processes over heterogeneous-in-nature, inter-related and massive data repositories.

A significant issue in dealing with OLAP data processing and querying is represented by the so-called *curse of dimensionality problem* [1], which, briefly, consists in the fact that when size and number of dimensions of the target data cube increase, multidimensional data cannot be accessed and queried efficiently. Starting from this practical evidence, several *Approximate OLAP Query Answering* techniques have been proposed during the last years, with alternate fortune. The main idea of these techniques consists in computing *compressed representations* of input data cubes in order to evaluate time-consuming OLAP queries against them, thus obtaining *approximate answers*. Despite compression introduces some approximation in the retrieved answers, it has been demonstrated [2] that fast and approximate answers are perfectly suitable to OLAP analysis goals, whereas exact and time-consuming answers introduce excessive computational overheads that, in general, are very often incompatible with the requirements posed by an online computation for decision making, as a very large number of tuples must be accessed in order to retrieve the desired exact answers.

This research proposes an excerpt of a ten-year-long history of OLAP data cube compression techniques, by particularly focusing on three major results: (*i*) $\Delta - Syn$ [4], an *analytical synopsis data structure* that introduces a polynomial approximation technique for OLAP data cubes; (*ii*) K_{LSA} [3], which further extends the $\Delta - Syn$ proposal in order to provide *accuracy control* over compressed OLAP data cubes; (*iii*) $\mathcal{LCS} - Hist$ [5], a *histogram-based complex methodology* for compressing massive-in-size high-dimensional OLAP data cubes. In the following, we provide an overview on these OLAP data cube compression techniques.

$\Delta - Syn$: Analytical Synopsis Data Structures Supporting Polynomial Approximation of OLAP Data Cubes [4] $\Delta - Syn$ is a synopsis data structure for OLAP data cubes that is based on an innovative *analytical interpretation* of multidimensional data and the well-known *Least Squares Approximation* (LSA), which provides support for approximate aggregate query answering in OLAP. According to the $\Delta - Syn$ methodology, the input data cube is interpreted as *a set of data rows*, to which appropriate *Discrete Impulsive Distributions* are associated. The final synopsis data structure is obtained by approximating the *Cumulative Distribution Functions* of such distributions with a set of *polynomial coefficients* provided by the LSA method, and storing these coefficients instead of the original data. This allows us to achieve a compact representation of the original data cube, being the size of the polynomial coefficient set is bounded by the storage space $\mathcal{B}$ available for housing the compressed data structure. In this research, an efficient algorithm that takes the input data cube $\mathcal{D}$ and the storage space $\mathcal{B}$, and builds $\Delta - Syn$ with low spatio-temporal complexity is proposed. OLAP queries are issued on the compressed representation using an optimized ad-hoc procedure, thus reducing the number of disk accesses needed to retrieve (approximate) answers. As demonstrated in [4], $\Delta - Syn$ provides good performance on both synthetic and real data cubes, even in comparison with other well-known compression techniques presented in literature, such as *histograms*, *wavelets* and *random sampling*.

K_{LSA}**: Accuracy Control Over Compressed Data Cubes for Quality-of-Answer OLAP Tools** [3] K_{LSA} is a state-of-the-art OLAP data cube compression technique that further extends the $\Delta - Syn$ proposal in order to provide *accuracy control* over compressed OLAP data cubes. K_{LSA} allows us to drive the compression process of data cubes in dependence on the accuracy required by external OLAP users/applications via determining the *degree of approximation* of final answers by means of meaningfully exploiting theoretical foundations offered by the LSA method. The main idea of the K_{LSA} proposal relies on two major assertions: (i) rigorously modeling and handling the degree of approximation of retrieved answers to OLAP queries over synopsis data structures; (ii) efficiently providing approximate answers having a desired accuracy, through setting the latter as an input parameter of the entire LSA-based compression process. This results in a novel *parametric LSA method* that meaningfully extends the baseline method and allows us to introduce the so-called *accuracy-aware LSA compression technique*. Given an input data distribution, the parametric LSA method is able of computing the *best* approximating function for this distribution in dependence on a *fixed* (i.e., required) accuracy. This baseline task is in turn exploited by the accuracy-aware LSA compression technique to achieve *accuracy-aware* compressed OLAP data cubes. A secondary-but-relevant contribution of the K_{LSA} proposal is represented by some effective optimizations of the above-sketched data cube compression technique that allow higher effectiveness and higher compression ratios to be achieved. Most importantly, K_{LSA} enables the design and the development of next-generation *Data Warehousing and OLAP Server Systems*, called *Quality-of-Answer (QoA) OLAP Tools*, which introduce an innovative paradigm according to which OLAP users/applications and DW servers implement an *application protocol* such that the final degree of approximation of retrieved answers is established by trading-off the required accuracy and the amount of storage space available for housing the compressed representation of the target data cube. A comprehensive experimental campaign on the K_{LSA} performance in compressing OLAP data cubes and supporting approximate query answering on so-compressed data cubes against several kinds of synthetic multidimensional data sets clearly demonstrates the superiority of K_{LSA} over significant similar techniques [3].

$\mathcal{LCS} - Hist$**: Scalable Histogram-based Approximation of Massive-In-Size High-Dimensional OLAP Data Cubes** [5] The $\mathcal{LCS} - Hist$ proposal introduces a *histogram-based complex methodology* for compressing massive-in-size high-dimensional OLAP data cubes whose main goal consists in overcoming actual *scalability limitations* of popular histogram-based data cube compression approaches. Indeed, classical histograms perform well on small-in-size low-dimensional data cubes whereas they do not scale satisfactorily on massive high-dimensional data cubes. For this reason, when the latter kind of data cubes are considered, we generally observe a significant performance degradation in both representing the input data domain and introducing low (query) errors in the retrieved approximate answers. To adequately face-off this drawback, the methodology underlying $\mathcal{LCS} - Hist$ defines an innovative data cube compression methodology that combines a *collection* of intelligent multidimensional data modeling and processing techniques: (i) $\mathcal{L}$*inear programming*, (ii) $\mathcal{C}$*onstrained partitions of multidimensional data domains*, and (iii) $\mathcal{S}$*imilarity metrics on one-dimensional histograms*. The main motivation of the novel vision carried out by $\mathcal{LCS} - Hist$ is the following. Since traditional histogram-based data cube compression techniques expose problematic limitations when applied to massive high-dimensional data cubes, combine intelligent multidimensional data modeling and processing techniques in order to obtain a final compressed data structure, the multidimensional histogram $\mathcal{LCS} - Hist$, that, although paying something in terms of computational overheads, allows us to achieve excellent performance in both representing the input data cube and efficiently supporting approximate query answering to resource-intensive OLAP queries against the compressed data structure. As proven in the experimental evaluation and analysis provided in [5], contrary to state-of-the-art histogram-based data cube compression techniques, $\mathcal{LCS} - Hist$ ensures high scalability and efficiency on massive high-dimensional data cubes, the most common kind of data cubes one can find in real-life OLAP applications.

References

1. Berchtold, S., Bhm, C., Kriegel, H.-P.: The pyramid-technique: Towards breaking the curse of dimensionality. In: Proceedings of the 1998 International Conference on Management of Data (SIGMOD 1998), pp. 142–153 (1998)
2. Cuzzocrea, A.: Overcoming limitations of approximate query answering in olap. In: Proceedings of the 9th International Symposium on Database Engineering and Applications (IDEAS 2005), pp. 200–209 (2005)
3. Cuzzocrea, A.: Accuracy control in compressed multidimensional data cubes for quality of answer-based olap tools. In: Proceedings of the 18th International Conference on Scientific and Statistical Database Management (SSDBM 2006), pp. 301–310 (2006)
4. Cuzzocrea, A.: Improving range-sum query evaluation on data cubes via polynomial approximation. Data & Knowledge Engineering 56(2), 85–121 (2006)
5. Cuzzocrea, A., Serafino, P.: LCS-hist: Taming massive high-dimensional data cube compression. In: Proceedings of the 12nd International Conference on Extending Database Technology (EDBT 2009), pp. 768–779 (2009)
6. Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, A., Pellow, F., Pirahesh, H.: Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. Data Mining and Knowledge Discovery 1(1), 29–53 (1997)

Author Index