Tai-hoon Kim Hojjat Adeli
Dominik Slezak Frode Eika Sandnes
Xiaofeng Song Kyo-il Chung
Kirk P. Arnett (Eds.)

LNCS 7105

Future Generation Information Technology

Third International Conference, FGIT 2011
in Conjunction with GDC 2011
Jeju Island, Korea, December 2011, Proceedings

Springer

Lecture Notes in Computer Science 7105

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

Tai-hoon Kim Hojjat Adeli Dominik Slezak
Frode Eika Sandnes Xiaofeng Song
Kyo-il Chung Kirk P. Arnett (Eds.)

Future Generation Information Technology

Third International Conference, FGIT 2011
in Conjunction with GDC 2011
Jeju Island, Korea, December 8-10, 2011
Proceedings

Springer

Volume Editors

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

Hojjat Adeli
The Ohio State University, Columbus, OH, USA
E-mail: adeli.1@osu.edu

Dominik Slezak
Infobright, Toronto, ON, Canada
E-mail: dominik.slezak@infobright.com

Frode Eika Sandnes
Oslo University College, Norway
E-mail: frode-eika.sandnes@hioa.no

Xiaofeng Song
Nanjing University of Aeronautics and Astronautics, Nanjing, China
E-mail: xfsong@nuaa.edu.cn

Kyo-il Chung
Electronics and Communications Research Institute (ETRI), Daejeon, Korea
E-mail: kyoil@etri.re.kr

Kirk P. Arnett
Mississippi State University, Oktibbeha, MS, USA
E-mail: kpal@msstate.edu

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-27141-0 e-ISBN 978-3-642-27142-7
DOI 10.1007/978-3-642-27142-7
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: Applied for

CR Subject Classification (1998): H.4, I.2, H.3, C.2, D.2, H.5

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

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

The Third International Mega-Conference on Future-Generation Information Technology (FGIT 2011) was held during December 8–10, 2011 in Jeju Grand Hotel, Jeju Island, Korea. This was composed of the following conferences: ASEA 2011, BSBT 2011, CA 2011, CES3 2011, DRBC 2011, DTA 2011, EL 2011, FGCN 2011, GDC 2011, MulGraB 2011, SecTech 2011, SIP 2011 and UNESST 2011.

The goal of this conference is to bring together researchers from academia and industry as well as practitioners to share ideas, problems and solutions relating to the multifaceted aspects of information technology.

We would like to express our gratitude to all of the authors of submitted papers and to all attendees for their contributions and participation.

We acknowledge the great effort of all the Chairs and the members of Advisory Boards and Program Committees of the above-listed event. Special thanks go to SERSC (Science and Engineering Research Support Society) for supporting this conference.

We are grateful in particular to the speakers who kindly accepted our invitation and, in this way, helped to meet the objectives of the conference.

December 2011 Chairs of FGIT 2011

Preface

We would like to welcome you to the proceedings of the 2011 Third International Mega-Conference on Future-Generation Information Technology (FGIT 2011) held during December 8–10, 2011, at Jeju Grand Hotel, Jeju Island, Korea.

FGIT is the most comprehensive conference focused on the various aspects of advances in information technology. The conference provides a chance for academic and industry professionals to discuss recent progress in the area of information technology.

We would like to acknowledge the great effort of the FGIT 2011 Chairs, Committees, International Advisory Board, as well as all the organizations and individuals who supported the idea of publishing this volume of proceedings, including the SERSC and Springer.

We are grateful to the following keynote, plenary and tutorial speakers who kindly accepted our invitation: Hsiao-Hwa Chen (National Cheng Kung University, Taiwan), Hamid R. Arabnia (University of Georgia, USA), Sabah Mohammed (Lakehead University, Canada), Ruay-Shiung Chang (National Dong Hwa University, Taiwan), Lei Li (Hosei University, Japan), Tadashi Dohi (Hiroshima University, Japan), Carlos Ramos (Polytechnic of Porto, Portugal), Marcin Szczuka (The University of Warsaw, Poland), Gerald Schaefer (Loughborough University, UK), Jinan Fiaidhi (Lakehead University, Canada) and Peter L. Stanchev (Kettering University, USA), Shusaku Tsumoto (Shimane University, Japan), Jemal H. Abawajy (Deakin University, Australia).

We would like to express our gratitude to all of the authors and reviewers of submitted papers and to all attendees, for their contributions and participation, and for believing in the need to continue this undertaking in the future.

Last but not the least, we give special thanks to Ronnie D. Caytiles and Yvette E. Gelogo of the graduate school of Hannam University, who contributed to the editing process of this volume with great passion.

This work was supported by the Korean Federation of Science and Technology Societies Grant funded by the Korean Government.

December 2011

Tai-hoon Kim
Hojjat Adeli
Dominik Slezak
Frode Eika Sandnes
Xiaofeng Song
Kyo-il Chung
Kirk P. Arnett

Organization

Honorary Chair

Young-hoon Lee	Hannam University, Korea

Steering Co-chairs

Tai-hoon Kim	GVSA and University of Tasmania, Australia
Wai-chi Fang	National Chiao Tung University, Taiwan

International Advisory Board

Haeng-kon Kim	Catholic University of Daegu, Korea
Tughrul Arslan	Engineering and Electronics, Edinburgh University, UK
Adrian Stoica	NASA Jet Propulsion Laboratory, USA
Yanchun Zhang	Victoria University, Australia
Stephen S. Yau	Chair, 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 Slezak	Infobright, Poland and Canada

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
Aboul Ella Hassanien	Cairo University, Egypt

Publication Co-chairs

Bongen Gu	Chungju National University, Korea
Younghwan Bang	KITECH, Korea

Table of Contents

Capturing Behavior of Medical Staff: A Similarity-Oriented Temporal Data Mining Approach*

Shusaku Tsumoto[1], Shoji Hirano[1], Haruko Iwata[1], and Yuko Tsumoto[1]

[1] Department of Medical Informatics, School of Medicine, Faculty of Medicine
Shimane University
89-1 Enya-cho Izumo 693-8501 Japan
{tsumoto,hirano,haruko23}@med.shimane-u.ac.jp
[2] Department of Fundamental Nursing, School of Nursing, Faculty of Medicine
Shimane University
89-1 Enya-cho Izumo 693-8501 Japan
tsumotoy@med.shimane-u.ac.jp

Abstract. This paper presents data mining results in which temporal behavior of global hospital activities are visualized. The results show that the reuse of stored data will give a powerful tool for hospital management and lead to improvement of hospital services.

Keywords: temporal data mining, visualization, clustering, hospital information system.

1 Introduction

Twenty years have passed since clinical data were stored electronically as a hospital information system. Stored data give all the histories of clinical activities in a hospital, including accounting information, laboratory data and electronic patient records. Due to the traceability of all the information, a hospital cannot function without the information system.

However, reuse of the stored data has not yet been discussed in details, except for laboratory data and accounting information to which OLAP methodologies are applied. Data mining approach just started ten years ago [3,4].

In this paper, we first propose a scheme for innovation of hospital services based on data mining. Then, based on this scheme, we applied data mining techniques to data extracted from hospital information systems. The results show several interesting results, which suggests that the reuse of stored data will give a powerful tool to improve the quality of hospital services.

The paper is organized as follows. Section 2 proposes a general framework on innovation of hospital services based on data mining. Section 3 briefly explains how hospital information system works, which is a background on this study.

* This research is supported by Grant-in-Aid for Scientific Research (B) 21300052 from Japan Society for the Promotion of Science (JSPS).

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 1–16, 2011.

Section 4 gives explanations on data preparation and mining process. Section 5 shows the results of visualization of hospital activities by using HIS data. Section 6 shows clustering-based analysis of similarities between divisions. Section 7 applies trajectories mining technique to temporal analysis the number of orders. Finally, Section 9 concludes this paper.

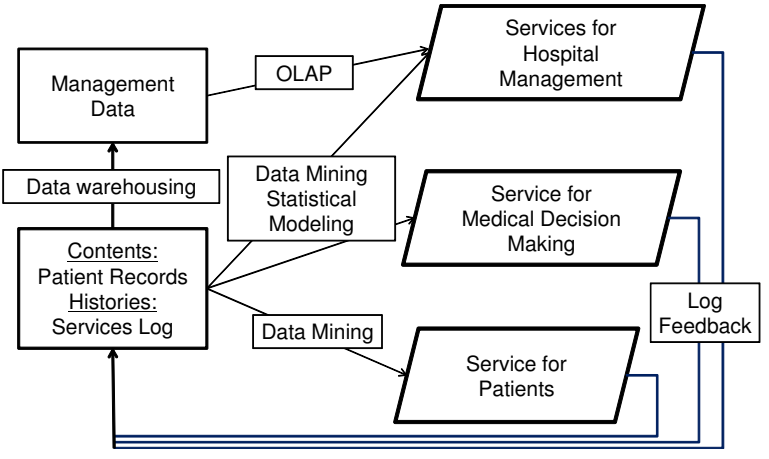


Fig. 1. Service-Oriented Hospital Management

2 Our Goal: Data-Mining Based Hospital Services

Figure 1 shows our goal for hospital services, which consists of the following three layers of hospital management: services for hospital management, devices for medical staff and services for patients. Data mining in hospital information system plays a central role in achieving these layers.

The first layer is called services for patients. It supports the improvement of healthcare service delivery for patients. This is a fundamental level of healthcare services in which medical staff directly gives medical services to the patients. Patient records and other results of clinical examinations support the quality of this service. The second layer is called services for medical staff. It supports decision making of medical practitioner. Patient histories and clinical data are applied to data mining techniques which gives useful patterns for medical practice. Especially, detection of risk of patients, such as drug adverse effects or temporal status of chronic diseases will improve the qualities of medical services. The top

layer is called services for hospital management. This level is achieved by capturing global behavior of a hospital: the bridging between microscopic behavior of medical staff and macroscopic behavior of hospital is very important to deploy medical staff in an optimal way for improving performance of the hospital.

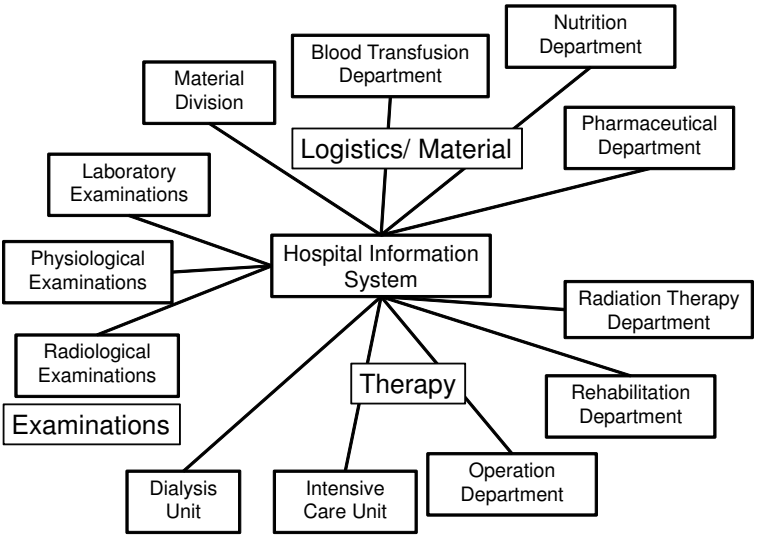


Fig. 2. Hospital Information System in Shimane University

3 Background

3.1 Hospital Information System: Cyberspace in Hospital

On the other hand, clinical information have been stored electronically as a hospital information system (HIS). The database stores all the data related with medical actions, including accounting information, laboratory examinations, and patient records described by medical staffs. Incident or accident reports are not exception: they are also stored in HIS as clinical databases. For example, Figure 2 shows the structure of the HIS in Shimane University Hospital. As shown in the figure, all the clinical inputs are shared through the network service in which medical staff can retrieve their information from their terminals [1,5].

Since all the clinical data are distributed stored and connected as a large-scale network, HIS can be viewed as a cyberspace in a hospital: all the results of clinical actions are stored as "histories". It is expected that similar techniques in data mining, web mining or network analysis can be applied to the data. Dealing with cyberspace in a hospital will give a new challenging problem in hospital

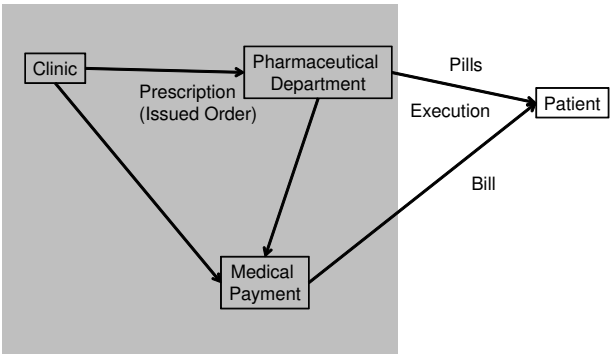


Fig. 3. Workflow of Prescription Order

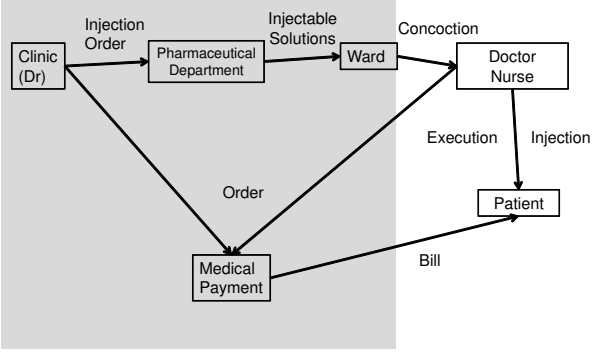


Fig. 4. Workflow of Injection Order

management in which spatiotemporal data mining, social network analysis and other new data mining methods may play central roles [8].

3.2 Basic Unit in HIS: Order

The basic unit in HIS is an "order", which is a kind of document or message which conveys an order from a medical practitioner to others. For example, prescription can be viewed as an order from a doctor to a pharmacist and an prescription order is executed as follows.

1. Outpatient Clinic
2. A prescription given from a doctor to a patient
3. The patient bring it to medical payment department
4. The patient bring it to pharmaceutical department

5. Execution of order in pharmacist office
6. Delivery of prescribed medication
7. Payment

The second to fourth steps can be viewed as information propagation: thus, if we transmit the prescription through the network, all the departments involved in this order can easily share the ordered information and execute the order immediately. This also means that all the results of the prescription process are stored in HIS.

Figure 3 depicts the workflow of prescription between doctors, pharmacologist, patients and pay desk. For comparison, Figure 4 shows that of injection.

These sharing and storing process, including histories of orders and their results, are automatically collected as a database: HIS can also be viewed as a cyberspace of medical orders.

4 Data Preparation and Analysis

4.1 DWH

Since data in hospital information systems are stored as histories of clinical actions, the raw data should be compiled to those accessible to data mining methods. Although this is usually called "data warehousing", medical data warehousing is different from conventional ones in the following three points. First, since hospital information system consists of distributed and heterogenous data sources. Second, temporal management is important for medical services, so summarization of data should include temporal information. Third, compilation with several levels of granularity is required. In this paper, we focus on the number of orders to capture temporal global characteristics of clinical activities, whose scheme is given as Figure 5. Here, data warehousing has two stages: first, we compile the data from heterogenous data sets with a given focus as the first DWH. Then, we split the primary DWH into two secondary DWHs: contents and histories. In this analysis, we focus on the latter DWH and we count the number of orders within a given temporal section. Data mining process is applied to the data sets generated from such obtained DWH.

4.2 Mining Process

We propose temporal data mining process, which consists of the following three steps, shown in Figure 6. We count temporal change of #orders per hour or per days in the second DWH. Then, since each order can be viewed as a temporal sequence, we compare these sequences by calculating similarities. Using similarities, clustering, multidimensional scaling (MDS) , and other similarity-based method are applied.

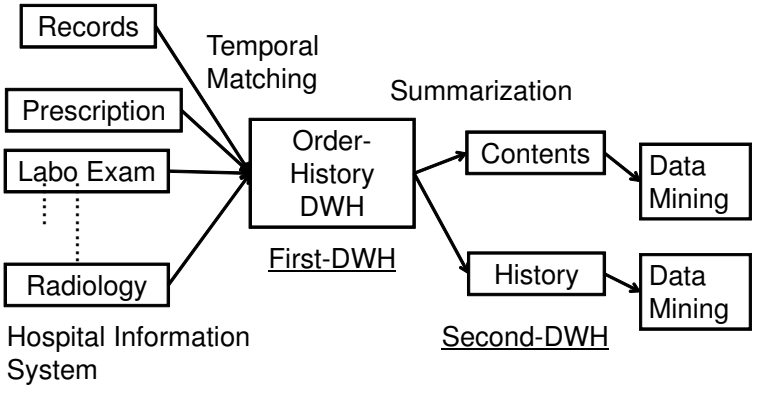


Fig. 5. Data warehousing

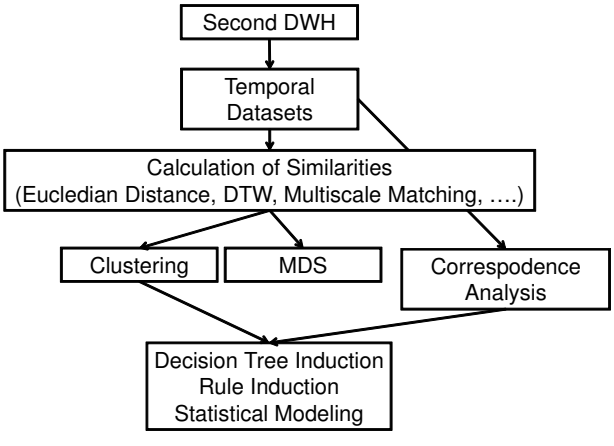


Fig. 6. Mining Process

4.3 Mining Methods

In this paper, the whole analysis is conducted by R2-13-1, except for trajectories mining. As for MDS and correspondence analysis, default methods were applied and for clustering, the ward method was applied. The library mvpart was used for decision tree mining. Multiscale matching was used for preprocessing for trajectories mining [6].

5 Visualizing Hospital Actions from Data

Let us show the primitive mining results of HIS. Table 1 shows the averaged number of each order during the same period. Although these values do not

Table 1. Averaged Number of Orders per day (Oct, 2006 to Feb, 2011)

	Outpatient			Ward		
	Average	Per Patient	Percentage	Average	Per Patient	Percentage
Prescription	403.010	0.457	11%	259.855	0.626	7%
Labo Exam	287.953	0.326	8%	221.373	0.533	6%
Phys Exam	78.595	0.089	2%	45.620	0.109	1%
Radiology	218.781	0.248	6%	56.352	0.135	2%
Operation			0%	12.853	0.031	0%
Transfusion	1.168	0.001	0%	11.089	0.027	0%
Meal			0%	186.293	0.448	5%
Pathology	21.909	0.025	1%	16.737	0.040	0%
Injection	84.414	0.096	2%	469.394	1.131	13%
Reservations	663.192	0.752	18%	77.465	0.186	2%
Documents	454.485	0.516	13%	156.966	0.378	5%
Nursery	11.087	0.0126	0%	846.334	2.0395	24%
Process	422.718	0.480	12%	99.0420	0.239	3%
Records	951.955	1.080	26%	963.017	2.321	28%
Rehabilitation	3.048	0.0035	0%	3.065	0.0074	0%
In/Out			0%	55.403	0.134	2%
Total	3602.315	4.088	100%	3480.856	8.388	100%
#Patients	881.209			414.972		

remove the effects of holidays, all the characteristics reflect those shown in Figure 2. Patient records and Nursing cares are the major part of orders (39%). Prescription, reservation of clinics, injection are top three orders in the hospital.

Table 2. Time Differences between Ordered and Performed Date (Oct, 2010)

	Average (Days)	Median (Days)
Prescription	0	0
Laboratory Exam.	28.55	14
Physiological Exam.	29.97	7
Radiology	29.22	9
Transfusion	8.897	6
Pathology	-0.003	0
Injection	9.799	2
Reservation	41.85	28
Nursery	1.397	0
Process	-0.0003 0	
Rehabilitation	0	0

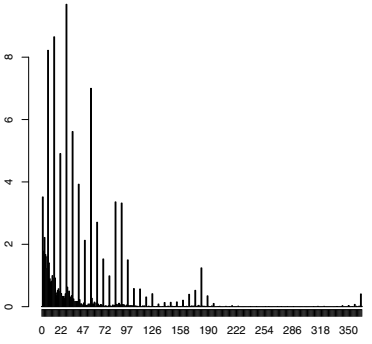


Fig. 7. Distribution of Differences between Executed and Issued Dates for Reservation

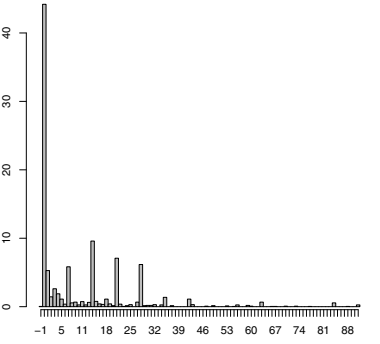


Fig. 8. Distribution of Differences between Executed and Issued Dates for Laboratory Examinations

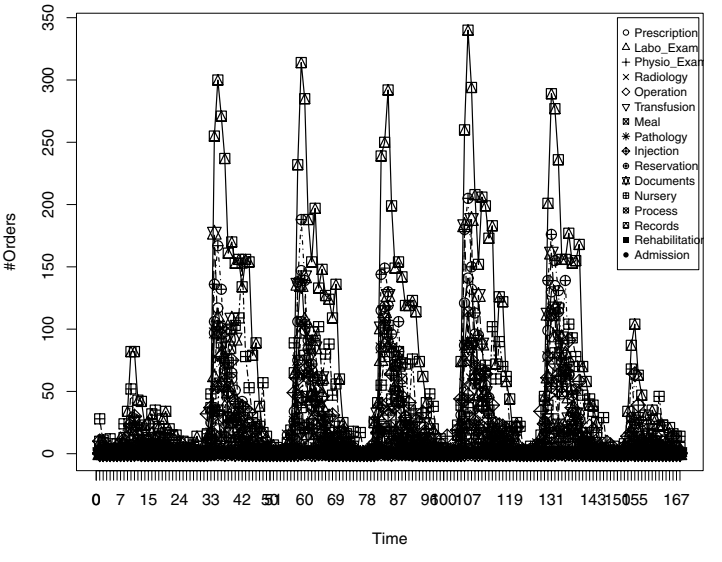


Fig. 9. Trends of Number of Orders (June 1 to 6, 2008)

Table 2 gives the statistics of time differences between ordered and performed date. For example, laboratory examinations are performed 28.55 days (averaged) and 14 days (median) after they are ordered.

Figure 7 shows the histogram (distribution) of time-difference on reservations. The peaks are given as 14, 21 , 28, 56, 90 and 180, which reflects the follow up period frequently used by clinicians.

Figure 8 shows the histogram (distribution) of time-difference on injections. The peaks are given as 7, 14, 21 and 28, which reflects the follow up period frequently used by clinicians.

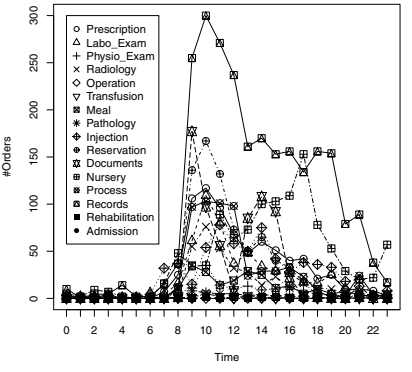


Fig. 10. Trends of Number of Orders (June 2, 2008)

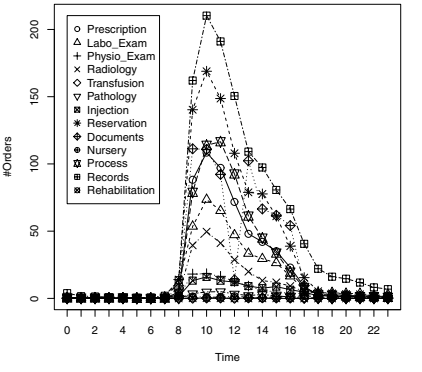


Fig. 11. Trends of Number of Orders of Outpatient Clinic (2010)

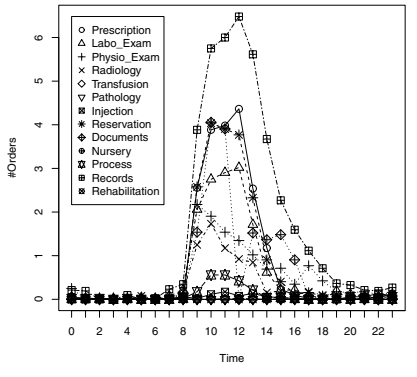


Fig. 12. Trends of Number of Orders of Cardiology (2010)

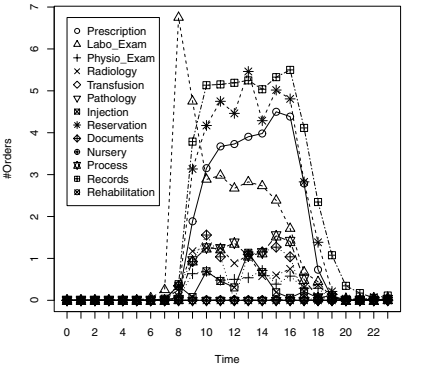


Fig. 13. Trends of Number of Orders of Rheumatology (2010)

5.1 Temporal Trend of #Orders

Although the above tables overview the total behavior of the hospital, we can also check the temporal trend of each order as shown in Figure 9 and 10. The former figure depicts the chronological overview of the number of each order from June 1 to 7, 2008, and the latter shows that of June 2, 2008. Vertical axes denote the averaged number of each order, classified by the type of orders. Horizonal axis give each time zone. The plots show the characteristics of each order. For example, the number of records of doctors has its peak in 11am, which corresponds to the peak of outpatient clinic, whose trend is very similar to reservation of outpatient clinic. The difference between these two orders is shown in 1pm to 5pm, which corresponds to the activities of wards.

The trends can capture the differences between division. Figures 12 and 13 show those in orders of outpatient clinics of cardiology and rheumatology on Tuesday. Compared with the total trends in outpatient clinic shown in Figure 11,

those of cardiology are much closer than those of rheumatology. These results show that we can measure and visualize the dynamics of clinical activities in the university hospital by exploratory methods. If we can detect some abnormalities different from the usual behavior in these measurements, this may give some knowledge about risks in the clinical activities. Thus, it is highly expected that data mining methods, especially spatiotemporal data mining techniques play crucial roles in analyzing data in hospital information system and understanding the dynamics of hospital [8,9].

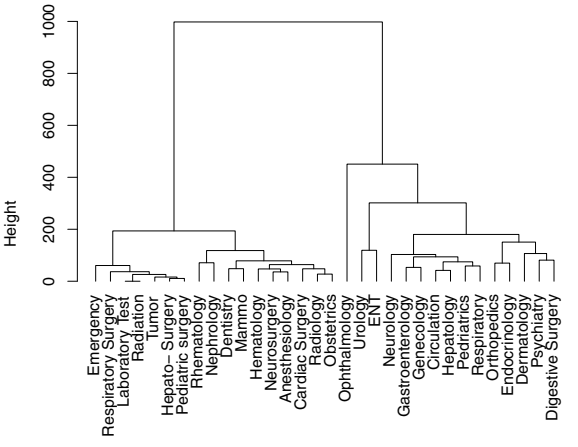


Fig. 14. Clustering of Chronological Patterns of Divisions

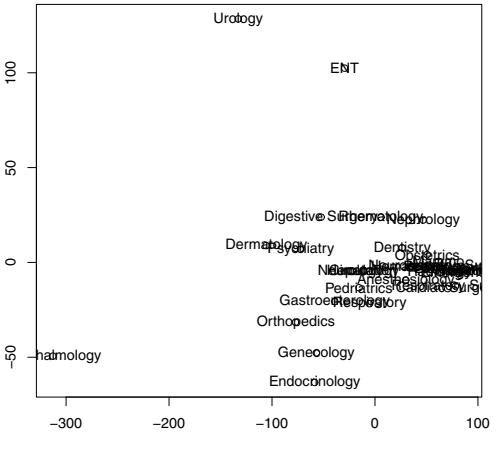


Fig. 15. MDS Results of Chronological Patterns of Divisions

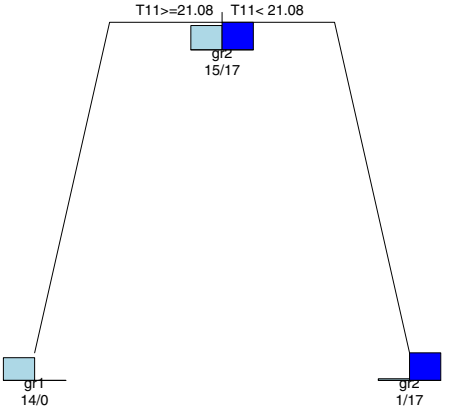


Fig. 16. Decision Tree induced by Chronological Patterns of # Total Orders

6 Clustering of Temporal Trends of Divisions

If we regard temporal trends as sequences for divisions in a hospital, we can classify divisions by using clustering methods.[1]

Figure 14 shows the grouping of divisions of hospitals by using Ward's method, in which the metrics are calculated from the chronological trends of the number of total orders in outpatient clinic. Figure 15 shows the results of multidimensional scaling, which gives a two-dimensional complementary view of similarities among divisions.

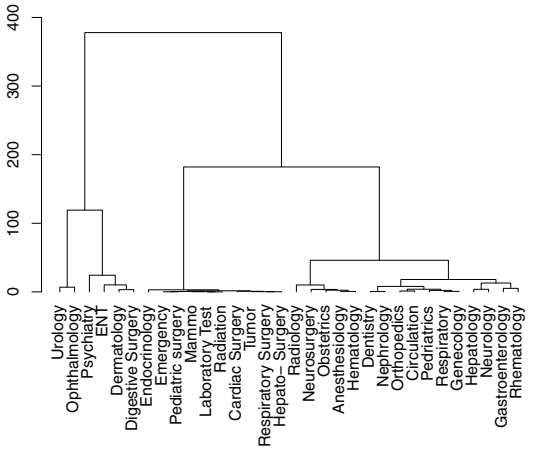


Fig. 17. Clustering of Division with T12

[1] Attributes may not be dependent, thus the usage of all the attributes will give us overfitted classification. Thus, feature selection should be considered: an approach to this problem is discussed in [7].

6.1 Decision Tree Induction

With the labels obtained, decision tree induction was applied to the data set. Figure 16 shows the result where only selected attribute is "T11", Tuesday 11pm. Thus, this time zone is the best attribute for classification of two major groups.

With this selected attribute, clustering method was refined as shown in Figure 17. Intuitively, this group is much better than the former one. Then, decision tree can be applied to data with newly generated groups. These results and repetitive process for temporal data mining is proposed in [7].

7 Analysis of Trajectory of #Orders

If we take two variables of each orders shown in Figure 10, then we can depict the trend of two attributes as a trajectory, as shown in Figures 18 and 19.

Tsumoto and Hirano proposes a clustering method of trajectories, which calculates dissimilarity measures via multiscale matching and apply clustering methods to trajectories by using the dissimilarities between trajectories [6]. By applying this method to the data shown in Section 5, the dendrogram shown in Figure 20 was obtained. An example of clusters are shown in Figure 21, which gives a pattern where orders are given both in wards and outpatient clinics. The other one gives a pattern where orders are provided mainly in the wards. A typical example in the first cluster is shown in Figure 18, while one in the second cluster is in Figure 19.

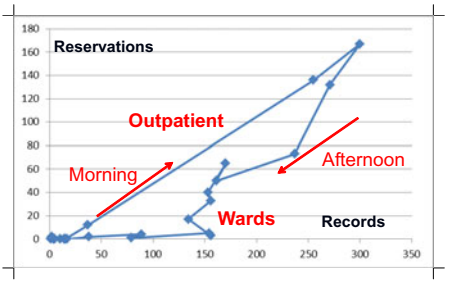


Fig. 18. Trajectory between #Reservations and #Records (June 2, 2008)

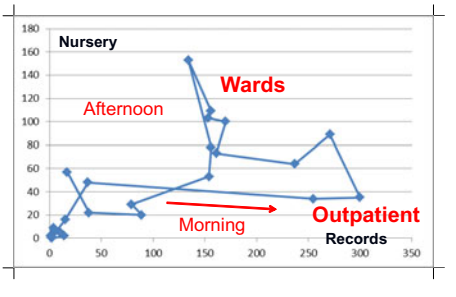


Fig. 19. Trajectory between #Nursery Orders and #Records (June 2, 2008)

The next step is to introduce three-dimensional trajectories mining proposed in [2], although selection of three variables plays an important role in making efficient classifications. It will be our future work to apply this method to the data.

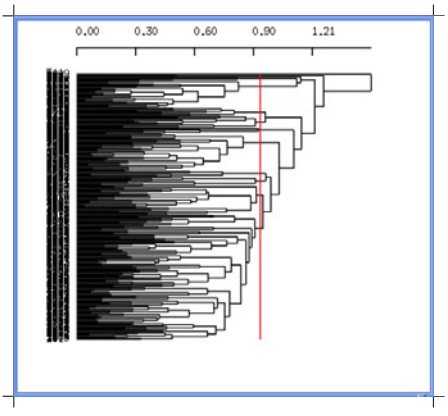


Fig. 20. Dendrogram of Trajectories (June 1 to 6, 2008)

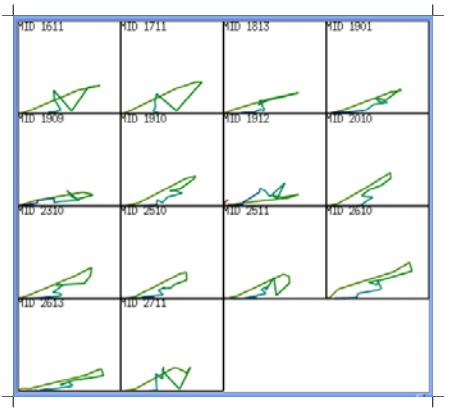


Fig. 21. Cluster No.1 (June 1 to 6, 2008)

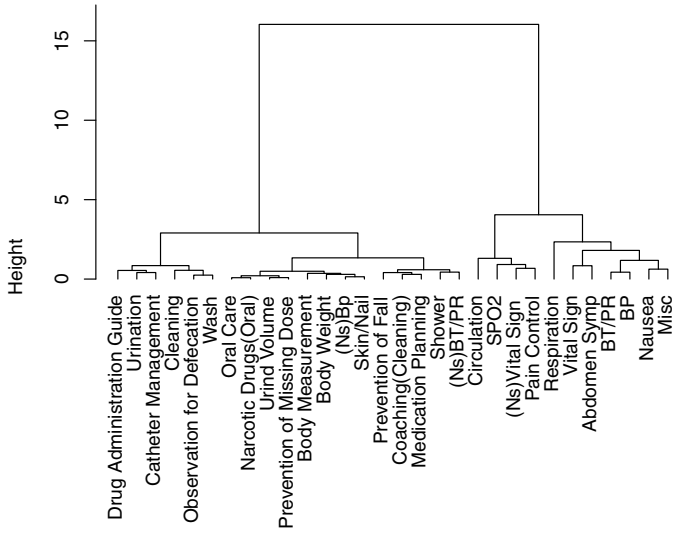


Fig. 22. Clustering Results of Nursing Orders (Lung Cancer)

8 Characterization of Nursing Orders

Finally, we focus on one disease, say lung cancer and count the nursing orders during the stay of each patient and regard chronological change of each order as a temporal sequence. Figures 22 to 24 show the results of clustering, MDS and correspondence analysis of nursing orders with respect to #orders.

Clustering results gave two major groups: one included the orders indispensable to this disease and the other included those which are rather specific to the status of each patient (Figure 28). MDS gave further classification of the

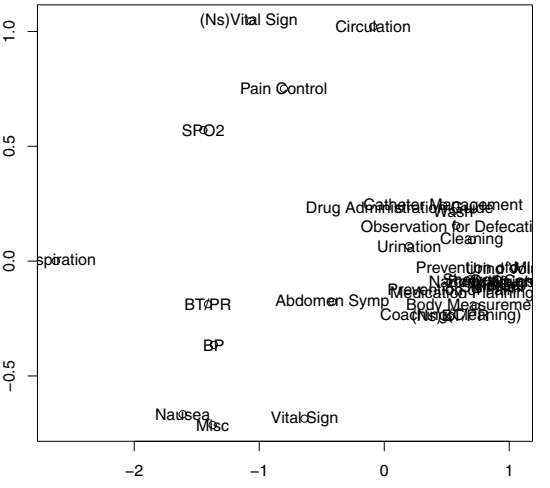


Fig. 23. MDS Results of Nursing Orders (Lung Cancer)

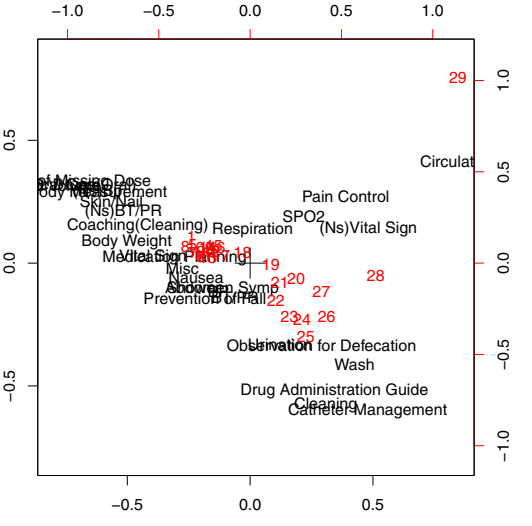


Fig. 24. Results of Correspondence Analysis of Nursing Orders (Lung Cancer)

first group into the following three subgroups: (1) core orders (Figure 25), (2) core and indispensable orders (Figure 27), both of which are not influenced by patients' status, and (3) core, indispensable orders which are influenced by their status (Figure 27). These results show that nursing orders can be classified by the proposed process, the deviation of this classification may give important information for a clinical action.

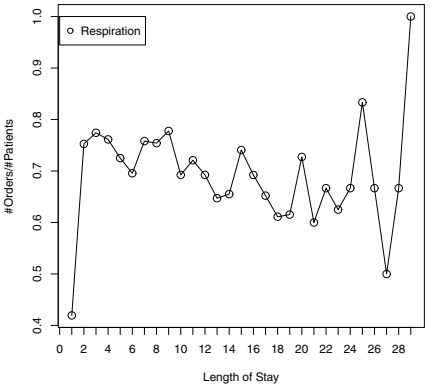


Fig. 25. The Main Nursing Orders

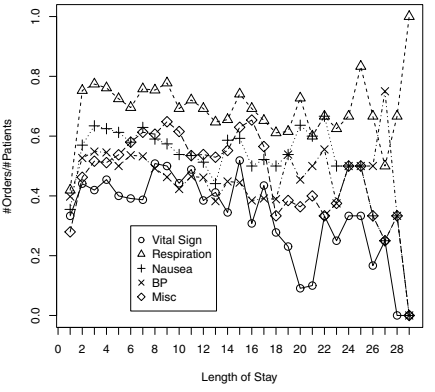


Fig. 26. Core orders which are not influenced by patients' status

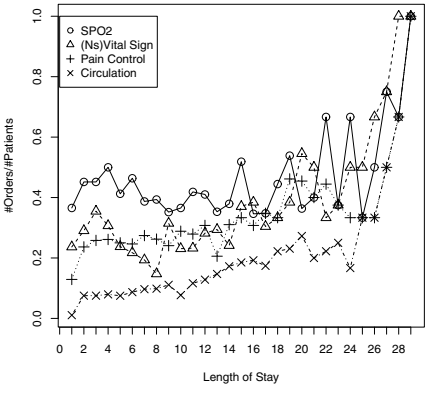


Fig. 27. Core Orders which may be influenced by patients' status

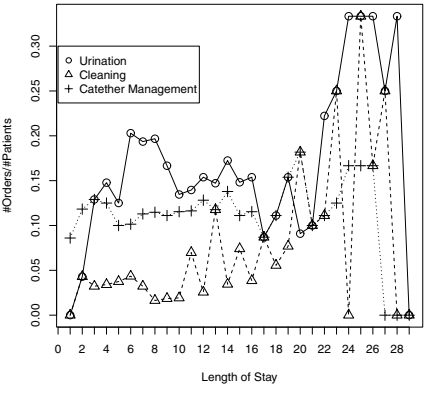


Fig. 28. Patient-specific Nursing Orders

9 Conclusions

In this paper, we propose a general framework on innovation of hospital services based on data mining. Then, we applied several data mining techniques to data extracted from HIS in order to capture the characteristics of clinical activities in hospital as follows. First, we shows the chronological overview of hospital activities: periodical behavior of the number of orders can be viewed as a "life-cycle" of a hospital. Secondly, we extracted a pattern of long-term follow up patients with respect to the number of orders by using HIS data. Section 6 shows the results of clustering analysis of divisions with respect to chronological change of total orders. Thirdly, we applied trajectories mining technique to temporal

analysis the number of orders. The results of clustering analysis gave two groups of clinical actions. The one was a pattern where orders are given both in wards and outpatient clinics. The other one was a pattern where orders are provided mainly in the wards. Finally, we applied similarity-based analysis to temporal trends of the numbers of nursing orders for lung cancer. The results showed that nursing orders are automatically classified into two major categories, "disease-specific" and "patient-specific" ones. Furthermore, the former one is classified into three subcategories, according to the temporal characteristics.

This paper is a preliminary approach to data mining hospital management towards a innovative process for hospital services. More detailed analysis will be reported in the near future.

References

1. Hanada, E., Tsumoto, S., Kobayashi, S.: A "Ubiquitous Environment" through Wireless Voice/Data Communication and a Fully Computerized Hospital Information System in a University Hospital. In: Takeda, H. (ed.) E-Health 2010. IFIP AICT, vol. 335, pp. 160–168. Springer, Heidelberg (2010)
2. Hirano, S., Tsumoto, S.: Multiscale comparison and clustering of three-dimensional trajectories based on curvature maxima. International Journal of Information Technology and Decision Making 9(6), 889–904 (2010)
3. Tsumoto, S.: Knowledge discovery in clinical databases and evaluation of discovered knowledge in outpatient clinic. Information Sciences (124), 125–137 (2000)
4. Tsumoto, S.: G5: Data mining in medicine. In: Kloesgen, W., Zytkow, J. (eds.) Handbook of Data Mining and Knowledge Discovery, pp. 798–807. Oxford University Press, Oxford (2001)
5. Tsumoto, S., Hirano, S.: Risk mining in medicine: Application of data mining to medical risk management. Fundam. Inform. 98(1), 107–121 (2010)
6. Tsumoto, S., Hirano, S.: Detection of risk factors using trajectory mining. J. Intell. Inf. Syst. 36(3), 403–425 (2011)
7. Tsumoto, S., Hirano, S., Tsumoto, Y.: Clustering-based analysis in hospital information systems. In: Proceedings of GrC 2011. IEEE Computer Society (2010)
8. Tsumoto, S., Hirano, S., Tsumoto, Y.: Towards data-oriented hospital services: Data mining-based hospital management. In: Fan, W., Hsu, W., Webb, G.I., Liu, B., Zhang, C., Gunopulos, D., Wu, X. (eds.) ICDM Workshops, pp. 1076–1083. IEEE Computer Society (2010)
9. Tsumoto, S., Hirano, S., Tsumoto, Y.: Information reuse in hospital information systems: A data mining approach. In: IRI, pp. 172–176. IEEE Systems, Man, and Cybernetics Society (2011)

AF or DF, and How to Configure an Optimal Mixed AF-DF Relay System?

Hsiao-Hwa Chen

Department of Engineering Science, National Cheng Kung University
1 Da-Hsueh Road, Tainan City, 70101 Taiwan
hshwchen@ieee.org

Channel fading due to multipath propagation in wireless communication systems causes a significant degradation in the received signal quality. Using various diversity reception techniques, we can mitigate signal degradation problem due to channel dispersion. However, due to the space limitations at mobile units, it is extremely difficult to have more than one antenna at the handsets. Thus, the cooperative relay technology provides us an excellent solution for high data-rate wireless transmission as required in the futuristic cellular and ad-hoc wireless communications systems.

Diversity technology plays an important role in the futuristic wireless communications. There are several different relay protocols proposed in the literature and they can be used to realize diversity signal reception. The most popular cooperation relay protocols [6] include amplify-and-forward (AF) and decode-and-forward (DF) [8] schemes. In addition, some other relay protocols were also suggested in the literature, including selection-and-forward and combined-and forward, which are however less commonly used than the AF and DF schemes. Many works have been done so far as an effort to study the performance of individual relay protocols (either AF or DF) for diversity signal reception. We have also seen a few papers published to investigate the outage and diversity order performance for a joint implementation with different relay protocols, where both AF and DF relay nodes may work together. However, the issues on their optimal configuration in terms of their diversity performance bounds have not been well investigated in the literature.

The authors in [10] and [2] examined the probability of DF protocol with an MRC receiver over Nakagami-m fading channels. However, the requirement of orthogonal relaying channels will lead to a loss in spectral efficiency. In contrast, DF selection cooperation [11,3] can avoid the reduction in spectral efficiency, but at a cost of implementation complexity.

The authors in [5] examined the performance of AF protocol with MRC receiver over Nakagami-m fading channels. However, to the best of our knowledge, no previous works [4,7,1,12,9] have thoroughly investigated the joint impact of the mixed AF and DF relay systems under independent Nakagami-m fading channels, and the exact analytical evaluation of the performance of the mixed AF and DF relay systems under independent Nakagami-m fading channels is still a widely open issue.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 17–18, 2011.

Our contributions in this work are two-fold, which can be summarized as follows: 1) we propose a framework to design a mixed AF and DF relay system which can give the optimal configuration under a specific channel condition better than either the only AF relay system or the only DF system; and 2) we derive the lower and upper bounds of the outage probability of a mixed AF and DF relay system based on MRC diversity combining with the help of order statistics.

We conclude that in a sense of maximized diversity order, a mixed AF and DF relay system with a similar number of AF and DF relays outperforms that with a largely different numbers of AF and DF relays, if the total number of relays is fixed. This conclusion gives an extremely useful guidance to the optimal implementation of a mixed AF and DF relay system for future wireless communication applications.

References

1. Cover, T.M., Thomas, J.A.: Element of Information Theory, 2nd edn. John Wiley & Sons, Inc. (2006)
2. Datsikas, C.K., Sagias, N.C., Lazarakis, F.I., Tombras, G.S.: Outage analysis of decode-and-forward relaying over Nakagami-m fading channels. IEEE Signal Processing Letters 15, 41–44 (2008)
3. Duong, T.Q., Bao, V.N.Q., Zepernick, H.J.: On the performance of selection decode-and-forward relay networks over Nakagami-m fading channels. IEEE Communications Letters 13(3), 172–174 (2009)
4. Gradshteyn, I.S., Ryzhik, I.M.: Table of Integrals, Series, and Products, 6th edn. Academic, New York (2000)
5. Ikki, S., Ahmed, M.H.: Performance analysis of cooperative diversity wireless networks over Nakagami-m fading channel. IEEE Communications Letters 11(4), 334–336 (2007)
6. Laneman, J.N., Tse, D.N.C., Wornell, G.W.: Distributed space-time-coded protocols for exploiting cooperative diversity in wireless networks. IEEE Trans. on Inform. Theory 49(10), 2415–2425 (2003)
7. Nakagami, M.: The m-distribution -A general formula of intensity distribution of rapid fading. In: Hoffman, W.G. (ed.) Statistical Methods in Radio Wave Propagation, pp. 3–36. Pergamon, Oxford (1960)
8. Savazzi, S., Spagnolini, U.: Cooperative fading regions for decode and forward relaying. IEEE Transactions on Information Theory 54(11) (2008)
9. Stark, H., Woods, J.W.: Probability and Random Processes with Applications to Signal Processing, 3rd edn. Prentice-Hall, Inc. (2002)
10. Savazzi, H.A.S., Smith, P.J., Armstrong, J.: Outage probability of cooperative relay networks in Nakagami-m fading channels. IEEE Communications Letters 10(12), 834–836 (2006)
11. Yan, K., Jiang, J., Wang, Y.G., Liu, H.T.: Outage probability of selection cooperation with MRC in Nakagami-m fading channels. IEEE Signal Processing Letters 16(12), 1031–1034 (2009)
12. Zhao, Y., Adve, R., Lim, T.J.: Outage probability at arbitrary SNR with cooperative diversity. IEEE Communications Letters 9(8), 700–702 (2005)

Cyber-Physical Intelligence in the Context of Power Systems

Carlos Ramos, Zita Vale, and Luiz Faria

GECAD – Knowledge Engineering and Decision Support Group / Institute of Engineering – Polytechnic of Porto, Portugal
{csr,zav,lef}@isep.ipp.pt

Abstract. Cyber-Physical Intelligence is a new concept integrating Cyber-Physical Systems and Intelligent Systems. The paradigm is centered in incorporating intelligent behavior in cyber-physical systems, until now too oriented to the operational technological aspects. In this paper we will describe the use of Cyber-Physical Intelligence in the context of Power Systems, namely in the use of Intelligent SCADA (Supervisory Control and Data Acquisition) systems at different levels of the Power System, from the Power Generation, Transmission, and Distribution Control Centers till the customers houses.

Keywords: Cyber-physical systems, Intelligent Systems, Power Systems, SCADA systems.

1 Introduction

Cyber-Physical Systems (CPS) can be defined as computing systems interacting with physical processes [1,2]. Physical systems are distributed in the real-world. Thus, it is expected that CPS inherits the distributed nature of Physical Systems too. For this reason CPS are conceived as networks of interacting elements instead of standalone devices. Information and Communication technologies are deeply embedded in CPS, allowing the interaction with physical processes to add new capabilities to physical systems. CPS must be dependable, secure, safe, efficient, and operate in real-time. They must also be scalable, cost-effective and adaptive. The integration of computational and physical processes exhibits a complex behavior that cannot be analyzed by the computational or physical sciences alone. These systems also transcend traditional computer-controlled systems because of their scale, dependence on man-machine interaction and their rich communication infrastructure that is enabled by the Internet. CPS range from very small to large-scale. Power Systems are referred as one of the most interesting examples of large-scale CPS applications, for example for achieving a blackout-free electricity generation and distribution system, for building smart grids, or for the optimization of energy consumption.

This paper is dedicated to the integration of Intelligence in Cyber-Physical Systems. We will illustrate this with Power Systems. Power Networks' critical physical infrastructure depends crucially on SCADA (Supervisory Control and Data Acquisition) and DCS (Digital Control Systems) for sensing, monitoring, gathering,

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 19–29, 2011.

and controlling distributed physical infrastructures. Power Systems Control Centers are the place where all the information of SCADA and DCS arrive, and Control Centre Operators' must handle the huge amount of data and information arriving from these systems, namely when we are in the presence of critical incidents. Now Cyber-Physical Intelligence is emerging as an important sub-area. The concept of Intelligent SCADA, with decentralized, flexible, and intelligent behavior, being dynamically adaptive to the context of the Power System is appearing.

Cyber-Physical Intelligence is a concept adequate to deal with Power Systems Smart Grids, Distributed Generation, mainly based on renewable sources, and Electricity Markets. We will present the project CITOPSY (Cyber-Ambient Intelligent Training of Operators in Power Systems Control Centres), in which the concept of Cyber-Physical Intelligence is being experimented.

In this paper we will describe how the concept of Intelligent SCADA [3] can be extended to other parts of the Power System, including the houses of the final customers.

2 Artificial Intelligence in Power Systems

Traditionally, Power Systems were conceived to provide electrical energy under security conditions. Guaranteeing sustainable development is a huge challenge for Power Systems. This requires a significant increasing in Distributed Generation, mainly based on renewable sources. However, this leads to a system that is much more complex to control, since we have many more Power Generation plants, and the generation is more unpredictable than before, due to the difficulty in forecasting the energy production in some renewable sources (e.g. wind and photovoltaic).

Intelligent Systems have been widely used in Power Systems. In [4] it is analyzed the application of Artificial Intelligence (AI) techniques to Power and Energy Systems. The main problems types where AI is used are the following:

- Alarm Processing, Diagnosis and Restoration - Expert systems and other knowledge-based systems have been used to address alarm processing, diagnosis and restoration support at dispatch and control center level but also at substation and generation plant level. When data concerning a large number of incidents, covering a large set of foreseen cases, is available, an artificial neural network can be designed and trained with good results. The Power industry performs the training of the Control Centre operators using simulators, some experiences have been done with Intelligent Tutoring Systems;
- Forecasting - The prevision of demand and of power generated by generation plants is very important for power systems. Neural Networks have been extensively used here. They have been also used in price forecasting for Electricity Markets, namely combined with Fuzzy Logic;
- Security Assessment - Security assessment evaluates the power system ability to face a set of contingencies, in static and dynamic situations. Pattern recognition methods and artificial neural networks have been used for security assessment;

- Planning and Scheduling – These techniques are necessary in electrical networks and generation expansion planning and in several operation problems like unit commitment, optimal dispatch, hydro thermal coordination, network reconfiguration and maintenance scheduling. The techniques more used to address these problems are computational intelligence and bio-inspired techniques, inspired in nature and animal behavior, including: artificial neural networks, fuzzy systems and genetic algorithms, simulated annealing, tabu search, swarm intelligence and ant colony;
- Energy markets – While Power Systems traditionally takes into account the technical nature of the problem, Energy Markets join an economical perspective to the Power Industry. Multi-Agent System, Machine Learning / Data Mining, and Game Theory techniques are used often in Energy Markets.

The introduction of Electricity Markets shows us the fragility of Power Systems infrastructures. Several severe incidents, including blackouts, occurred (e.g. the 14th August 2003 Blackout in USA, the 4th October 2006 quasi-blackout affecting 9 European countries, and the 10th November 2009 Brazilian blackout affecting 70 million people without electricity in 18 Brazilian states, including major southern cities São Paulo and Rio de Janeiro). However, the problem is more human than technical, since some of these incidents were caused or had increased consequences due to operators' mistakes, namely at the level of Power Systems Control Centers (CC)[5]. There is a trend to the degradation of this situation with the increasing in the complexity of Power Systems due to the augmenting number of Renewable source installations in the near future.

Thus, training Power Systems Control Centres' Operators for this new reality is a critical goal. For this purpose we are developing CITOPSY (Cyber-Ambient Intelligent Training of Operators in Power Systems CC) project [6]. The project considers the following emerging and advanced paradigms:

- Cyber-Physical Systems (CPS), since Power Systems networks, strongly based on SCADA systems, that are expected to evolve for Intelligent SCADA, are typically pointed as a reference example of CPS [1,2];
- Ambient Intelligence (AmI) [7,8,9,10], since we claim that CC are a very good example of environment where AmI makes sense;
- and Intelligent Tutoring System (ITS), an advanced technology that is now achieving the maturity to be used in real-world applications, and that is specially adequate for training, namely when combined with simulators. Here we have some experience in the development of ITS for Power Systems incident analysis and diagnosis, power restoration, according to an integrated view [11,12].

3 Intelligence at the Control Center Level

The main goal of this work is to supply CC operators with an intelligent environment, fully integrated with the SCADA system, and to provide the CC room with training

abilities. In order to accomplish this goal, we will proceed with the integration of an ITS for the diagnosis and restoration of Power Systems in a CC room.

3.1 Tutoring Environment Architecture

Figure 1 shows this tutoring environment architecture, composed of three complex systems: a CPS module responsible for the acquisition and treatment of the network's physical data; an AmI system (Interaction Manager) that allows the trainee's immersion in the control room environment; and a tutoring system responsible for the pedagogical process. The tutoring system involves two main areas: one devoted to the training of fault diagnosis skills and another dedicated to the training of power system restoration techniques.

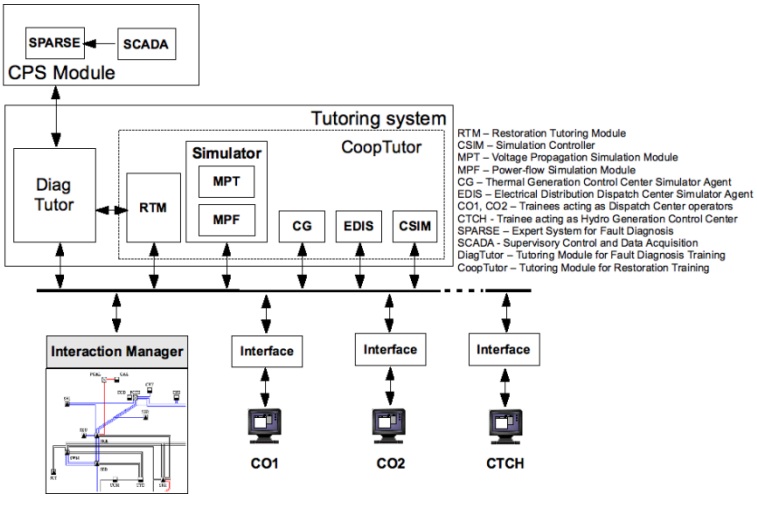


Fig. 1. Tutoring environment architecture

The selection of the adequate established restoration procedure strongly depends on the correct identification of the Power System operation state. Therefore, the identification of the incidents or set of incidents occurring in the transmission network is of utmost relevance in order to establish the current Power System operation state. Thus the proposed training framework divides operator's training in to distinct stages. The first one is intended to give operators with competence needed to get incident diagnosis. After that, operators are able to use CoopTutor to train their skills to manage the restoration procedures.

3.2 Tutoring Module for Fault Diagnosis Training

During the analysis of lists of alarm messages, CC operators must have in mind the group of messages that describes each type of fault. The same group of messages can show up in the reports of different types of faults. So CC operators have to analyze the arrival of additional information, whose presence or absence determines the final diagnosis.

Operators have to deal with uncertain, incomplete and inconsistent information, due to data loss or errors occurred in the data gathering system.

The interaction between the trainee and the tutor is performed through prediction tables where the operator selects a set of premises and the corresponding conclusion. The premises represent events (SCADA messages), temporal constraints between events or previous conclusions.

DiagTutor does not require the operator's reasoning to follow a predefined set of steps, as in other implementations of the model tracing technique [13]. In order to evaluate this reasoning, the tutor will compare the prediction tables' content with the specific situation model. This model is obtained by matching the domain model with the inference undertaken by SPARSE expert system [5]. This process is used to: identify the errors revealing operator's misconceptions; provide assistance on each problem solving action, if needed; monitor the trainee knowledge evolution; and provide learning opportunities for the trainee to reach mastery. In the area of ITS's this goals has been achieved through the use of cognitive tutors [14].

Indeed, the process used to evaluate the trainee's reasoning is based on the application of pattern-matching algorithms. Similar approaches with the same propose are used in other ITS's, such as in the TAO ITS [15], an ITS designed to provide tactical action officer students at US navy with practice-based and individualized instruction.

3.3 Tutoring Module for Restoration Training

The management of a power system involves several distinct entities, responsible for different parts of the network. The power system restoration needs a close coordination between generation, transmission and distribution personnel and their actions should be based on a careful planning and guided by adequate strategies [16].

In the specific case of the Portuguese transmission network, four main entities can be identified: the National Dispatch Center (CC), responsible for the energy management and for the thermal generation; the Operational Centre (CO), controlling the transmission network; the Hydroelectric Control Centers (CTCH), responsible for the remote control of hydroelectric power plants and the Distribution Dispatch (EDIS), controlling the distribution network.

The power restoration process is conducted by these entities in such a way that the parts of the grid they are responsible for will be slowly led to their normal state, by performing the actions specified in detailed operating procedures and fulfilling the

requirements defined in previously established protocols. This process requires frequent negotiation between entities, agreement on common goals to be achieved, and synchronization of the separate action plans on well-defined moments.

Several agents personify the four entities that are present in the power system restoration process. This multi-agent approach was chosen because it is the most natural way of translating the real-life roles and the split of domain knowledge and performed functions that can be witnessed in the actual power system. Several entities responsible for separate parts of the whole task must interact in a cooperative way towards the fulfillment of the same global purpose. Agents' technology has been considered well suited to domains where the data is split by distinct entities physically or logically and which must interact with one another to pursue a common goal [17].

In this section we describe how we developed a training environment able to deal adequately with the training of the procedures, plans and strategies of the power system restoration, using what may be called lightweight, limited scope simulation techniques. This environment's purpose is to make available to the trainees all the knowledge accumulated during years of network operation, translated into detailed power system restoration plans and strategies, in an expedite and flexible way. The embedded knowledge about procedures, plans and strategies should be easily revisable, any time that new field tests, post-incident analysis or simulations supply new data.

These agents can be seen as virtual entities that possess knowledge about the domain. As real operators, they have tasks assigned to them, goals to be achieved and beliefs about the network status and others agents' activity. They work asynchronously, performing their duties simultaneously and synchronizing their activities only when this need arises. Therefore, the system needs a facilitator (simulator in Fig. 1) that supervises the process, ensuring that the simulation is coherent and convincing.

In our system, the trainee can choose to play any of the available roles, namely the CO and the CC ones, leaving to the tutor the responsibility of simulating the other participants.

The ITS architecture was planned in order that future upgrades of the involved entities or the inclusion of new agents to be simple.

This tutoring module is able to train individual operators as if they were in a team, surrounded by virtual "operators", but is also capable of dealing with the interaction between several trainees engaged in a cooperative process. It provides specialized agents to fulfill the roles of the missing operators and, at the same time, monitors the cooperative work, stepping in when a serious imbalance is detected. The tutor can be used as a distance learning tool, with several operators being trained at different locations.

4 Intelligence at Customers House Level

Nowadays, the evolution of energy resources used in Power System (PS) and the roles of the involved players require new management approaches. Decisions concerning energy resource management are being increasingly decentralized and consumers are progressively gaining a central role in the power system efficiency [3]. In this context, house consumptions must be intelligently managed. Consumers with Demand Response (DR) contracts should manage their consumption according to their preferences and opportunities presented by DR programs [18,19].

A real-time load management based on artificial intelligence techniques seems to be the most effective way to achieve high efficiency levels while respecting consumers' requirements and preferences [20]. In order to fully attain the aimed goals, the load management system should be able to learn from consumers' actions and routines, which vary from one consumer to another, from season-to-season.

The GECAD Intelligent Energy Systems Laboratory (LASIE), located in the Institute of Engineering – Polytechnic of Porto (ISEP/IPP) includes an intelligent SCADA house simulation for which a SCADA system supervises and controls both domestic electricity generation and consumption. The generation system includes two photovoltaic panels (one fixed and one tracking), two wind turbines and one fuel-cell. Different loads are available, including variable loads (induction motors and fluorescent lamps) and discrete loads (lamps, heating and divers electric appliances) [20,21].

The proposed method for load management defines a consumption limit for each period, according to the existent consumption profile. However, several events that can occur may change the normal consumption profile, that requires considering a different consumption limit. This can be caused by:

- Fluctuation of consumers' owned generation - To adjust the consumption to the generation in each instant;
- System Operator Indication - Normally used to reduce the consumption in peak periods or in the case of unexpected outages. Requires direct load control demand response contracts;
- Energy price in real time – In order to keep the power system stability (maintaining consumption and generation balance), the system operator can change the energy price in real time waiting for the consumption response.

Whenever the consumption is higher than the defined consumption limit, it is required that SHIM (SCADA House Intelligent Management) intelligently reduces the consumption through load curtailment or reduction [20]. Figure 2 presents the implemented methodology.

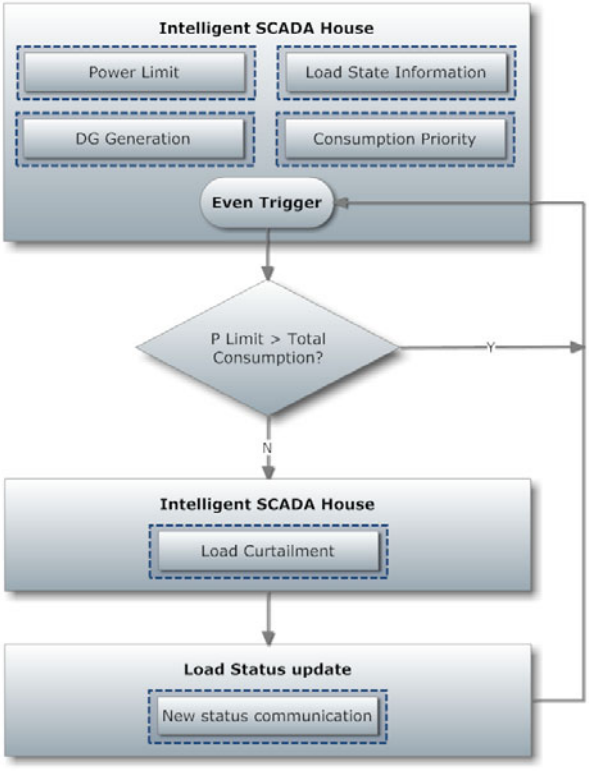


Fig. 2. Tutoring environment architecture

SHIM system presented in this paper considers the preferences defined in users' profiles. This means that the system needs to learn through experience so that it can automatically update users' profiles. The ability of learning is essential to obtain a system capable of adapting itself over the time without the intervention of the client or external entity.

The user profile can change throughout the system life. The users may have different routines in different week days or in different seasons. One of the most significant differences in the daily routines is between working days and weekend. In SHIM the working days and weekends are processed independently.

However, the users can change their routines along the system lifetime requiring the system to have a learning mechanism. SHIM machine learning mechanism is based on artificial neural networks.

In SHIM the learning process may change the preference factor of each variable (concerning the electrical appliances) to adjust their importance over time. SHIM learning module has been implemented in MATLAB.

Figure 3 shows SHIM's architecture. SHIM interacts with the user through an interface that allows commanding the electrical equipment, monitoring the consumption, and defining the consumption limits.

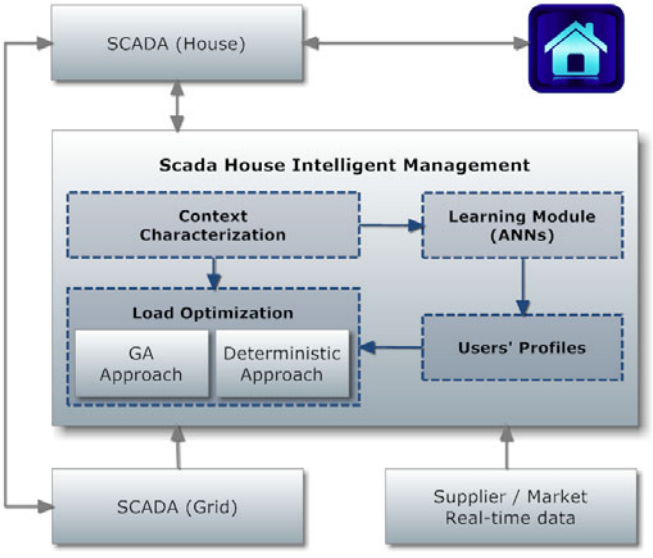


Fig. 3. Tutoring environment architecture

The interaction between the grid SCADA and the house is important when there are DR contracts that allow the grid operator to limit the house consumption. The communication between SHIM and the grid SCADA can be made with Power Line Communications (PLC), Global System Mobile (GSM) or Ethernet. Other important information is the supplier/market real time data. These data is available in smart meters and allow SHIM to optimize the use of some equipment such as electric boilers and electric vehicles charge / discharge.

In each instant the system undertakes the context characterization and this information is used in the learning process to adjust the user´s profile. At the same time, if the consumption is higher than the consumption limit, the load optimization process runs.

5 Conclusions

Cyber-Physical Systems need to be enhanced with Intelligent Systems features in order to allow intelligence at the different levels of distributed systems. This paper described how Power Systems can be improved with Intelligent SCADA. Two different levels were considered: power transmission in Power Systems Control Centers, and costumers houses. Several AI technologies are used, namely Intelligent Tutoring Systems, Expert Systems, Neural Networks, Genetic Algorithms, and Multi-Agent Systems.

Acknowledgements. The authors would like to acknowledge the Portuguese Science and technology Foundation (FCT), and COMPETE Program for their support to R&D Projects and GECAD Research Unit.

References

1. Lee, E.: Cyber Physical Systems: Design Challenges, University of California, Berkeley Technical Report No. UCB/EECS-2008-8
2. CPS Steering Group, Cyber-Physical Systems Executive Summary (2008), http://varma.ece.cmu.edu/Summit/CPS-Executive-Summary.pdf
3. Vale, Z., Morais, H., Silva, M., Ramos, C.: Towards a future SCADA. In: 2009 IEEE Power and Energy Society General Meeting General Meeting, Calgary, Alberta, Canada (2009)
4. Vale, Z.: Intelligent Power Systems. Wiley Encyclopedia of Computer Science and Engineering. John Wiley & Sons, Inc. (2008)
5. Vale, Z., Machado e Moura, A., Fernanda Fernandes, M., Marques, A., Rosado, C., Ramos, C.: SPARSE: An Intelligent Alarm Processor and Operator Assistant for Portuguese Substations Control Centers. IEEE Expert - Intelligent Systems and Their Application 12(3), 86–93 (1997)
6. Faria, L., Silva, A., Ramos, C., Vale, Z., Marques, A.: Intelligent Training in Control Centres Based on an Ambient Intelligence Paradigm. In: García-Pedrajas, N., Herrera, F., Fyfe, C., Benítez, J.M., Ali, M. (eds.) IEA/AIE 2010. LNCS, vol. 6096, pp. 143–153. Springer, Heidelberg (2010)
7. ISTAG, Strategic Orientations & Priorities for IST in FP6, European Commission Report (2002)
8. Ramos, C., Augusto, J.C., Shapiro, D.: Ambient Intelligence: the next step for AI. IEEE Intelligent Systems Magazine 23(2), 15–18 (2008)
9. Augusto, J.C., McCullagh, P.: Ambient Intelligence: Concepts and Applications. International Journal on Computer Science and Information Systems 4(1), 1–28 (2007)
10. Ramos, C.: Ambient Intelligence Environments. In: Rabuñal, J., Dorado, J., Sierra, A. (eds.) Encyclopedia of Artificial Intelligence, pp. 92–98. Information Science Reference (2009), ISBN 978-1-59904-849-9
11. Ramos, C., Frasson, C., Ramachandran, S.: Introduction to the Special Issue on Real World Applications of Intelligent Tutoring Systems. IEEE Transactions on Learning Technologies 2(2), 62–63 (2009), doi:10.1109/TLT.2009.16
12. Faria, L., Silva, A., Vale, Z., Marques, A.: Training Control Centres' Operators in Incident Diagnosis and Power Restoration Using Intelligent Tutoring Systems. IEEE Transactions on Learning Technologies (2009)
13. Anderson, J., Corbett, A., Koedinger, K., Pelletier, R.: Cognitive Tutors: Lessons Learned. The Journal of the Learning Sciences 4(2), 167–207 (1995)
14. Aleven, V., Koedinger, K.R.: An effective meta-cognitive strategy: learning by doing and explaining with a computer-based Cognitive Tutor. Cognitive Science 26(2), 147–179 (2002)
15. Stottler, R., Panichas, S.: A New Generation of Tactical Action Officer Intelligent Tutoring System (ITS). In: Proc. Industry/Interservice, Training, Simulation and Education Conference, I/ITSEC (2006)
16. Sforna, M., Bertanza, V.: Restoration Testing and Training in Italian ISO. IEEE Transactions on Power Systems 17(4) (November 2002)

17. Jennings, N., Wooldridge, M.: Applying agent technology. Applied Artificial Intelligence: An International Journal 9(4), 351–361 (1995)
18. Weihao, H., et al.: Optimal Load Response to Time-of-Use Power Price for Demand Side Management in Denmark. In: 2010 Asia-Pacific Power and Energy Engineering Conference (APPEEC), Chengdu, China (2010)
19. Faria, P., Vale, Z., Ferreira, J.: DemSi – A Demand Response Simulator in the context of intensive use of Distributed Generation. In: Proc. 2010 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2010), Istanbul (October 2010)
20. Fernandes, F., Sousa, T., Silva, M., Morais, H., Vale, Z., Faria, P.: Genetic Algorithm Methodology applied to Intelligent House Control. In: Symposium on Computational Intelligence Applications in Smart Grid (CIASG) at the IEEE SSCI 2011 (IEEE Symposium Series on Computational Intelligence), Paris, France, April 11-15 (2011)
21. Fernandes, F., Sousa, T., Faria, P., Silva, M., Morais, H., Vale, Z.: Intelligent SCADA for Load Control. In: IEEE International Conference on Systems, Man and Cybernetics, SMC 2010, Istanbul, Turkey, October 12-15 (2010)

Interactive Navigation of Image Collections*

Gerald Schaefer

Department of Computer Science
Loughborough University
Loughborough, U.K.
gerald.schaefer@ieee.org

Abstract. Image databases are growing at a rapid rate and hence efficient and effective techniques to manage these vast repositories are highly sought after. Image database navigation systems provide an interesting alternative to retrieval based approaches, and in this paper we show how image browsers can be used for interactive exploration of large image collections based on the principle that visually similar images are located close to each other thus helping user navigation, and that large datasets are handled through a hierarchical approach.

1 Introduction

Image databases are growing at a rapid rate and may contain millions of images [1]. Hence efficient and effective techniques to manage these vast repositories are highly sought after. Unfortunately, only a small percentage of images get annotated in practice [2] which in turn has significant implications for query systems as no textual information can be used for retrieval. However, content-based retrieval methods [3,4,5,6,7,8] can be employed which extract various image features (describing e.g. colour, texture or shape properties) as descriptors and allow retrieval of images based on derived visual similarity. These features can also be used for image database navigation systems which present an interesting alternative to image retrieval approaches [9,10,11,12,13].

The idea of image database navigation systems is to provide a visualisation of a complete image collection [14] together with browsing tools for an interactive exploration of the database [15]. Approaches to image browsing can in general be divided into three categories [14]:-

- *mapping-based techniques*, which typically employ dimensionality reduction algorithms to map the high-dimensional feature space to a low-dimensional space for visualisation;
- *clustering-based techniques*, which group visually similar images together using clustering algorithms;

* The author wishes to acknowledge William Plant, the main developer of the desktop-based browser, Matthew Tallyn and Daniel Felton for porting the system to the iOS environment, and Matthew Fox who added multi-touch functionality to the application.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 30–36, 2011.

– *graph-based techniques*, which use graph structures where graph nodes correspond to images and edges indicate relationships (e.g., visual similarity) between images.

In this paper, we focus on one of the image browsing systems we have developed in our lab, namely the Honeycomb Image Browser [16]. In a sense, our approach can be seen as a combination of mapping-based and clustering-based image browsing. However, to significantly reduce the computational complexity, we neither employ dimensionality reduction techniques nor clustering algorithms. Images are grouped, based on visually similarity, in the visualisation space which is arranged on a hexagonal lattice. To allow access also to large image repositories, the database is organised in a hierarchical fashion where the user can select any of the representative images shown and expand the relevant image cluster. We have also ported the system to operate on mobile devices and on large multi-touch screens.

2 Honeycomb Image Browser

The Honeycomb Image Browser [16] is based on several ideas originally employed by the Hue Sphere Image browser [17,18,19,20,21], one of our earlier image database navigation systems. We arrange images by visual similarity so that similar images are located close to each other in visualisation space. As features for expressing similarity we employ very simple colour descriptors, namely the median colour expressed in HSV colour space. This has the advantage that

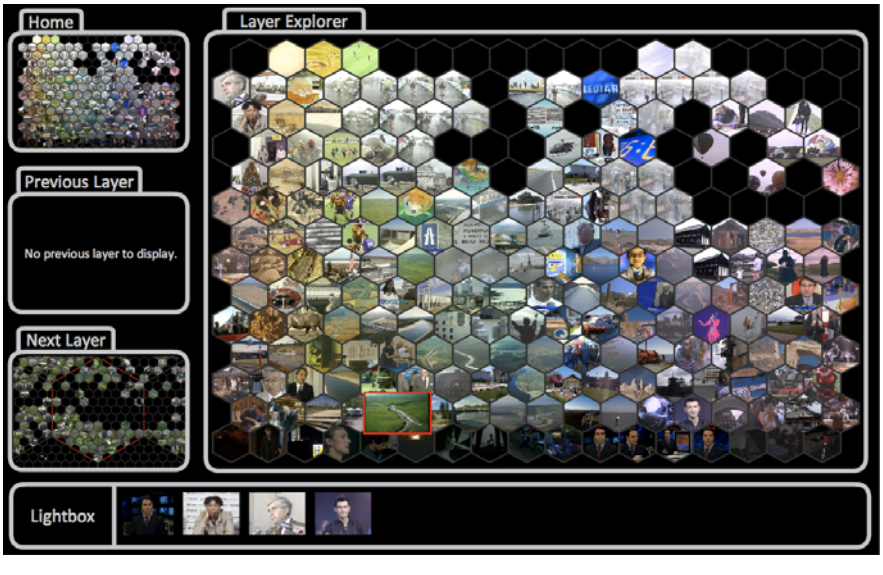


Fig. 1. Image database visualised in the Honeycomb Image Browser

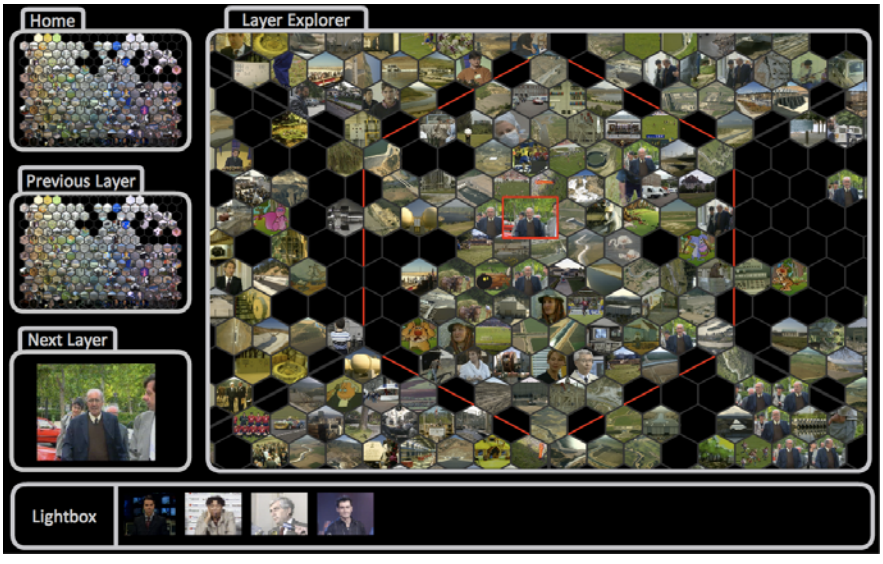


Fig. 2. Honeycomb Image Browser after navigating into the next layer of the visualisation

it greatly reduces the overall computational complexity of our approach, as on the one hand the features themselves can be calculated extremely fast, while on the other hand no computationally intensive dimensionality reduction technique is necessary. Rather, the location of each image is derived directly from the hue and value (brightness) co-ordinates.

The visualisation space is then divided using a regular hexagonal lattice. This means that images cannot overlap nor occlude each other, which in turn has been shown to lead to an improved browsing experience [22]. Each image in the database will fall into exactly one hexagon, and we can hence make use of the advantages of clustering-based methods without actually having to employ a clustering technique, which again makes our approach decisively less demanding in terms of computational load. Large databases are handled by employing a hierarchical approach to visualising and browsing images. If more than one image falls into a specific cell, a representative image (that closest to the centre of the cell) will be shown in the browser, while the user has the possibility to open that image cluster and hence navigate to the next level of the browsing hierarchy. Here, the colour space is again divided into (now smaller) hexagons and the same principles as on the root layer are employed. To ensure more cells are filled, two spreading strategies are applied where the first one moves images from cells to neighbouring empty cells, while the second method spreads out images that are very similar to its surrounding neighbours.

The browsing operations the user can apply include panning and optical zoom (very much in the style of document viewers) on every layer of the visualisation as well as vertical browsing [15] which allows the user to navigate the hierarchical navigation structure by expanding (or returning to) specific image clusters. In addition, users can place images of interest in a lightbox for further processing or export. In Figures 1 and 2 we show screenshots of the browsing application run on the MPEG-7 Common Colour Dataset [23].

3 Mobile Browsing Interface

As more and more images are being taken with and stored on mobile phones (rather than with digital cameras / on desktops computers or websites), we have ported the Honeycomb Image Browser to mobile devices such as smartphones. In particular, we have created a version of the browser for the iOS environment[1] so that it can be used on devices such as the iPhone and iPod.

Clearly, the reduction of screen space presents a challenge, and therefore the resolution of the hexagonal lattice had to be reduced. We also removed the root, previous layer and preview layer displays to allow maximum usage of the screen area. User interaction in the mobile browser is based on various touch gestures (rather than mouse or keyboard events as in the desktop version); panning and optical zooming are implemented in the expected manner (e.g., optical zooming is triggered by 2-finger pinch gestures) while entering a lower level of the browsing hierarchy is performed upon a double-tap on the respective hexagon. Examples of a Mobile Honeycomb Image Browser session are shown in Figures 3 and 4.

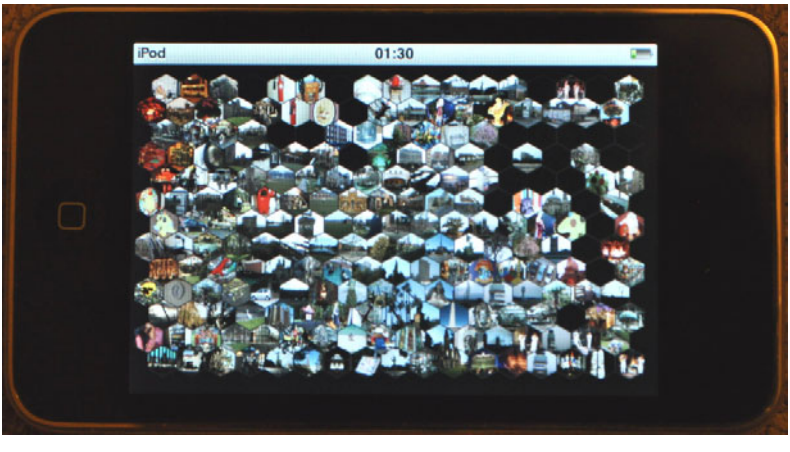


Fig. 3. Image database visualised in the Mobile Honeycomb Image Browser on an iPod

[1] http://developer.apple.com/devcenter/ios

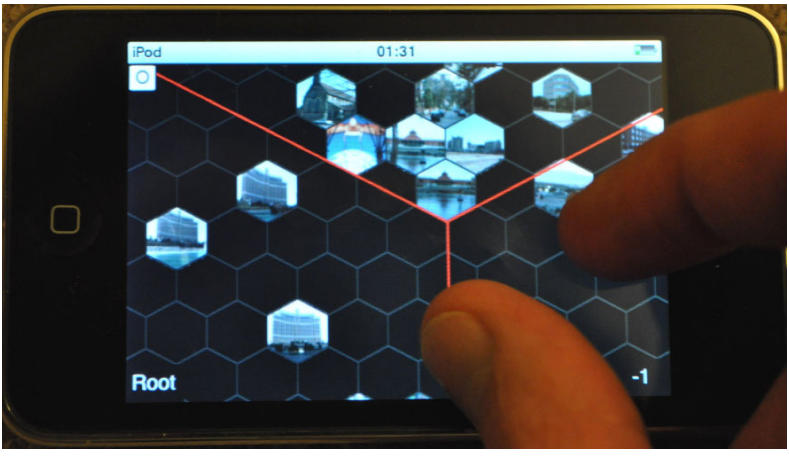


Fig. 4. Mobile Honeycomb Image Browser after navigating into the next layer of the browsing structure

4 Multi-touch Interface

While our port to smartphones as outlined above meant working on a relatively small screen size, we have also gone "in the opposite direction", and implemented the browser for a large multi-touch screen, a PQ Labs G^3 screen with a

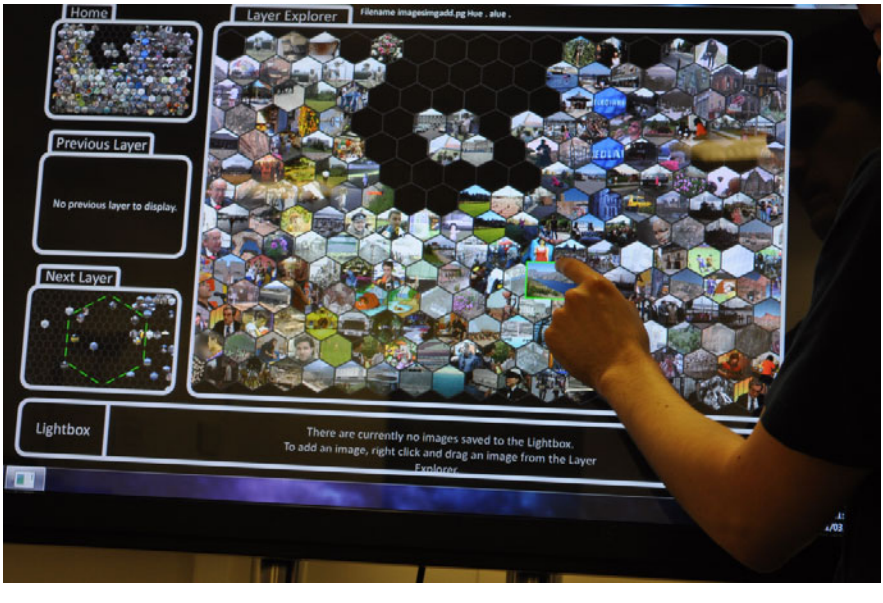


Fig. 5. Multi-touch Honeycomb Image Browser

resolution of 1600x1200 pixels. Similar to the mobile device version, we have added touch gestures for panning, zooming, opening image clusters and other browsing operations. Figure 5 shows the multi-touch application of the generated browser.

5 Conclusions

Image database navigation tools present an interesting alternative to direct retrieval approaches, in particular as they allow true user interaction with an image collection. In this paper, we have summarised the Honeycomb Image Browser as a good example of such tools, and have also presented recent ports of the system to mobile devices and to large multi-touch screens.

References

1. Osman, T., Thakker, D., Schaefer, G., Lakin, P.: An integrative semantic framework for image annotation and retrieval. In: IEEE/WIC/ACM International Conference on Web Intelligence, pp. 366–373 (2007)
2. Rodden, K.: Evaluating Similarity-Based Visualisations as Interfaces for Image Browsing. PhD thesis, University of Cambridge Computer Laboratory (2001)
3. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Analysis and Machine Intelligence 22, 1249–1380 (2000)
4. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys 40, 1–60 (2008)
5. Schaefer, G.: Search and retrieval of images by content. In: Int. Conference on Web Technologies and Internet Applications, pp. 5–8 (2011)
6. Schaefer, G.: Mining Image Databases by Content. In: Belhajjame, K. (ed.) BNCOD 2011. LNCS, vol. 7051, pp. 66–67. Springer, Heidelberg (2011)
7. Schaefer, G.: Content-Based Image Retrieval: Some Basics. In: Czachórski, T., Kozielski, S., Stańczyk, U. (eds.) Man-Machine Interactions 2. AISC, vol. 103, pp. 21–29. Springer, Heidelberg (2011)
8. Schaefer, G.: Content-Based Image Retrieval: Advanced Topics. In: Czachórski, T., Kozielski, S., Stańczyk, U. (eds.) Man-Machine Interactions 2. AISC, vol. 103, pp. 31–37. Springer, Heidelberg (2011)
9. Plant, W., Schaefer, G.: Visualisation and Browsing of Image Databases. In: Lin, W., Tao, D., Kacprzyk, J., Li, Z., Izquierdo, E., Wang, H. (eds.) Multimedia Analysis, Processing and Communications. SCI, vol. 346, pp. 3–57. Springer, Heidelberg (2011)
10. Schaefer, G.: Content-based retrieval from image databases: colour, compression, and browsing. In: Int. Conference on Information Retrieval and Knowledge Management, pp. 5–10 (2010)
11. Schaefer, G.: Visualisation and browsing of large image repositories. In: 10th Int. Conference on Information (2010)
12. Schaefer, G.: Image browsers effective and efficient tools for managing large image collections. In: 2nd Int. Conference on Multimedia Computing and Systems, pp. 1–3 (2011)

13. Schaefer, G.: Interactive Exploration of Image Collections. In: Burduk, R., Kurzyński, M., Woźniak, M., Żołnierek, A. (eds.) Computer Recognition Systems 4. AISC, vol. 95, pp. 229–238. Springer, Heidelberg (2011)
14. Plant, W., Schaefer, G.: Visualising image databases. In: IEEE Int. Workshop on Multimedia Signal Processing, pp. 1–6 (2009)
15. Plant, W., Schaefer, G.: Navigation and browsing of image databases. In: Int. Conference on Soft Computing and Pattern Recognition, pp. 750–755 (2009)
16. Plant, W., Schaefer, G.: Image retrieval on the honeycomb image browser. In: 17th IEEE Int. Conference on Image Processing, pp. 3161–3164 (2010)
17. Schaefer, G., Ruszala, S.: Image Database Navigation: A Globe-Al Approach. In: Bebis, G., Boyle, R., Koracin, D., Parvin, B. (eds.) ISVC 2005. LNCS, vol. 3804, pp. 279–286. Springer, Heidelberg (2005)
18. Schaefer, G., Ruszala, S.: Hierarchical Image Database Navigation on a Hue Sphere. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Remagnino, P., Nefian, A., Meenakshisundaram, G., Pascucci, V., Zara, J., Molineros, J., Theisel, H., Malzbender, T. (eds.) ISVC 2006. LNCS, vol. 4292, pp. 814–823. Springer, Heidelberg (2006)
19. Schaefer, G., Ruszala, S.: Effective and efficient browsing of image databases. Int. Journal of Imaging Systems and Technology 18, 137–145 (2008)
20. Schaefer, G.: A next generation browsing environment for large image repositories. Multimedia Tools and Applications 47, 105–120 (2010)
21. Schaefer, G., Stuttard, M.: An on-line tool for browsing large image repositories. In: Int. Conference on Information Retrieval and Knowledge Management, pp. 102–106 (2010)
22. Rodden, K., Basalaj, W., Sinclair, D., Wood, K.: Evaluating a visualisation of image similarity as a tool for image browsing. In: IEEE Symposium on Information Visualization, pp. 36–43 (1999)
23. Moving Picture Experts Group: Description of core experiments for MPEG-7 color/texture descriptors. Technical Report ISO/IEC JTC1/SC29/WG11/ N2929 (1999)

Environmental Diversity Techniques of Software Systems

Tadashi Dohi

Department of Information Engineering, Graduate School of Engineering
Hiroshima University, 1–4–1 Kagamiyama, Higashi-Hiroshima, 739–8527 Japan
dohi@rel.hiroshima-u.ac.jp

Several recent studies have reported that most outages in technical computer-based systems are due to software faults. Traditional methods in software engineering are fault avoidance/removal based on extensive testing/debugging, and fault tolerance based on design/data diversity. Since both of them are very expensive and unrealistic in common cases, the key challenge is how to provide highly dependable software with relatively cheaper cost. We introduce several environmental diversity techniques of software systems, and overview the typical examples involving checkpoint restart and software rejuvenation. Based on the author's own research results during a past decade, we discuss stochastic models to derive several checkpoint restart and software rejuvenation policies analytically in terms of the optimality under cost criteria.

First, we formulate the checkpoint placement problems [1,14,16,17], which can be characterized by optimization problems to derive the checkpoint sequence. Second, we introduce the concept of software aging and rejuvenation, and summarize several stochastic models to determine the optimal software ejuvenation policies [2,3,4,5,7,8,9,10,11,12,13,14,15,18,19]. Third, we concern intrusion tolerant systems and consider some control policies to improve the system availability and its related measures [6,20,21,22]. Finally, the present research trend and the open problems in future are also discussed.

References

1. Dohi, T., Kaio, N., Osaki, S.: The optimal age-dependent checkpoint strategy for a stochastic system subject to general failure mode. Journal of Mathematical Analysis and Applications 249, 80–94 (2000)
2. Dohi, T., Goseva-Popstojanova, K., Trivedi, K.S.: Estimating software rejuvenation schedule in high assurance systems. The Computer Journal 44(6), 473–485 (2001)
3. Dohi, T., Iwamoto, K., Okamura, H., Kaio, N.: Discrete availability models to rejuvenate a telecommunication billing application. IEICE Transactions on Communications (B) E86-B(10), 2931–2939 (2003)
4. Dohi, T., Suzuki, H., Trivedi, K.S.: Comparing software rejuvenation policies under different dependability measures. IEICE Transactions on Information and Systems (D) E87-D(8), 2078–2085 (2004)
5. Dohi, T., Suzuki, H., Osaki, S.: Transient cost analysis of non-Markovian software systems with rejuvenation. International Journal of Performability Engineering 2(3), 233–243 (2006)

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 37–38, 2011.

6. Dohi, T., Uemura, T.: An adaptive mode control algorithm of a scalable intrusion tolerant architecture. Journal of Computer and System Sciences (in press)
7. Eto, H., Dohi, T.: Determining the optimal software rejuvenation schedule via semi-Markov decision process. Journal of Computer Science 2(6), 528–534 (2006)
8. Iwamoto, K., Dohi, T., Kaio, N.: Estimating periodic software rejuvenation schedule in discrete operational circumstance. IEICE Transactions on Information and Systems (D) E91-D(1) (2008)
9. Okamura, H., Miyahara, S., Dohi, T., Osaki, S.: Performance evaluation of workload-based software rejuvenation scheme. IEICE Transactions on Information and Systems (D) E84-D(10), 1368–1375 (2001)
10. Okamura, H., Miyahara, S., Dohi, T.: Dependability analysis of a transactionbased multi server system with rejuvenation. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences (A) E86-A (8), 2081–2090 (2003)
11. Okamura, H., Fujio, H., Dohi, T.: Fine-grained shock models to rejuvenate software systems. IEICE Transactions on Information and Systems (D) E86-D(10), 2165–2171 (2003)
12. Okamura, H., Miyahara, S., Dohi, T.: Rejuvenating communication network system with burst arrival. IEICE Transactions on Communications (B) E88-B(12), 4498–4506 (2005)
13. Okamura, H., Iwamoto, K., Dohi, T.: A dynamic programming algorithm for software rejuvenation scheduling under distributed computation circumstance. Journal of Computer Science 2(6), 505–512 (2006)
14. Okamura, H., Iwamoto, K., Dohi, T.: A DP-based optimal checkpointing algorithm for real-time appications. International Journal of Reliability, Quality and Safety Engineering 13(4), 323–340 (2006)
15. Okamura, H., Dohi, T.: Comprehensive evaluation of aperiodic checkpointing and rejuvenation schemes in operational software system. Journal of Systems and Software 83, 1591–1604 (2010)
16. Ozaki, T., Dohi, T., Okamura, H., Kaio, N.: Distribution-free checkpoint placement algorithms based on min-max principle. IEEE Transactions on Dependable and Secure Computing 3(2), 130–140 (2006)
17. Ozaki, T., Dohi, T., Kaio, N.: Numerical computation algorithms for sequential checkpoint placement. Performance Evaluation 66, 311–326 (2009)
18. Rinsaka, K., Dohi, T.: Behavioral analysis of fault-torellant software systems with rejuvenation. IEICE Transactions on Information and Systems (D) E88-D(12), 2681–2690 (2005)
19. Suzuki, H., Dohi, T., Okamura, H.: Cost-effective analysis of periodic software rejuvenation policies for a telecommunication billing application. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences (A) E85-A (12), 2923–2932 (2002)
20. Uemura, T., Dohi, T.: Optimal security patch management policies maximizing system availability. Journal of Communications 5(1), 71–80 (2010)
21. Uemura, T., Dohi, T., Kaio, N.: Availability analysis of an intrusion tolerant distributed server system with preventive maintenance. IEEE Transactions on Reliability 59(1), 18–29 (2010)
22. Uemura, T., Dohi, T., Kaio, N.: Dependability analysis of a scalable intrusion tolerant architecture with two detection modes. Journal of Internet Technology 11(2), 289–298 (2010)

Multimedia Standards. History. State of Art*

Peter L. Stanchev[1,2]

[1] Institute of Mathematics and Informatics
Bulgarian Academy of Sciences, Sofia, Bulgaria
[2] Kettering University, Flint, USA
pstanche@kettering.edu

Abstract. The aim of this presentation is to review some of the standards, connected with multimedia and their metadata. We start with MPEG family and continue with Open Standards for Interactive TV. Efficient video-streaming is presented. Some standards for audiovisual metadata are outline. We finish with W3C standards.

Keywords: Multimedia Standards, MPEG, Interactive TV, Video-Stream Filtering, Audiovisual Media, W3C.

1 MPEG Standards

The Moving Picture Experts Group (MPEG) [6] is a working group of ISO/IEC in charge of the development of standards for coded representation of digital audio and video and related data. Since 1988 when it has been established, the group has produced standards that help the industry offer end users an ever more enjoyable digital media experience. The number of independent standards is more than 125. Follow is a list of standard families.

- MPEG-1, the standard for such products as Video CD and MP3 are based;
- MPEG-2, the standard for such products as Digital Television set top boxes and DVD are based;
- MPEG-4, the standard for multimedia for the fixed and mobile web;
- MPEG-7, the standard for description and search of audio and visual content;
- MPEG-21, the Multimedia Framework;
- MPEG-A, the standard for application-specific formats;
- MPEG-B, a collection of Systems specific standards;
- MPEG-C, a collection of Video specific standards;
- MPEG-D, a collection of Audio specific standards;
- MPEG-DASH, the standard for video streaming over the internet;
- MPEG-E, a standard (M3W) providing support to download and execution of multimedia applications;

* This work was supported in part by Open Access Infrastructure for Research in Europe (OpenAIRE) EU project, and the Bulgarian National Science Fund under the Project D002 308 "Automated Metadata Generating for e-Documents Specifications and Standards".

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 39–42, 2011.

- MPEG-H, a standard (HEVC) that will provide significantly increased video compression performance compared to AVC;
- MPEG-M, a standard (MXM) for packaging and reusability of MPEG technologies;
- MPEG-U, a standard for rich-media user interface;
- MPEG-V, a standard for interchange with virtual worlds.

In its 23 years of activity MPEG has developed an impressive portfolio of technologies that have created an industry worth several hundred billion USD. In the moment in the full development phase are the following standards:

- 3D Video Coding, the standard for coding 3D Visual information;
- High Efficiency Video Coding, the standard for a new frontier in video coding;
- MPEG-M 2nd edition, the standard for digital media ecosystems.

An example of MPEG-4 system is the ANIMATION system - a system for animation scene and contents creation, retrieval and display [9]. Semantic Video and Image Retrieval is analyzed in [10].

The upcoming Reconfigurable Video Coding (RVC) standard currently developed at MPEG supports the construction of video standards as libraries of coding tools. These libraries can be incrementally updated and extended, and the tools in them can be aggregated to form complete codecs using a streaming programming model, which preserves the inherent parallelism of the coding algorithm [4].

Depending on the specific characteristics of an image data set, some features can be more effective than others when performing similarity search. Starting from this observation, a technique that predicts the effectiveness of MPEG 7 image features based on a statistical analysis of the specific data sets in the Multimedia Content Management System was developed [8].

MPEG-V (ISO/IEC 23005) provide an architecture and associated information representations to enable the interoperability between virtual worlds, e.g., digital content provider of a virtual world, gaming, simulation, DVD, and with the real world, e.g., sensors, actuators, vision and rendering, robotics, support for independent living, social and welfare systems, banking, insurance, travel, real estate, rights management and many others [2].

2 Open Standards for Interactive TV

Levels of interactivity TV [7] can be described as: Level 1 - Basic TV, Level 2 - Call-In-TV, Level 3 - Parallel TV, Level 4 - Additive TV, Level 5 - Service on Demand, Communicative TV, and Level 7 - Fully Interactive TV. The Multimedia and Hypermedia Information Coding Expert Group (MHEG) [5], a subgroup of the International Organization for Standardization (ISO), published the MHEG standards. A digital video broadcaster filtering algorithm is presented in [1].

3 Standards in the Audiovisual Media

The main issues in the Audiovisual Media standards for audio visual media include [3]:

- EBU P Meta. The European Broadcasting Union (EBU) has defined P Meta as a metadata vocabulary for program exchange in the professional broadcast industry;
- Material Exchange Format. The Material Exchange Format (MXF) is a standard issued by Society of Motion Picture and Television Engineers (SMPTE), defining the specification of a file format for the wrapping and transport of essence and metadata in a single container.
- SMPTE Descriptive Metadata Scheme 1 (DMS-1), formerly known as Geneva Scheme uses metadata sets defined in the SMPTE Metadata Dictionary;
- SMPTE Metadata Dictionary (is a large thematically structured list of narrowly defined metadata elements, defined by a key, the size of the value and its semantics;
- Standard Media Exchange Framework (SMEF) the data model defined by the BBC to describe the metadata related to media items (media objects) and programs and parts thereof (editorial objects), down to the shot level;
- Controlled Vocabulary and Ontologies. Audiovisual content descriptions often contain references to semantic entities such as objects, events, states, places, and times.

4 W3C Standards

The World Wide Web Consortium (W3C) [11] develops technical specifications and guidelines through a process designed to maximize consensus about the content of a technical report, to ensure high technical and editorial quality, and to earn endorsement by W3C and the broader community. They include:

- Web Design and Applications involve the standards for building and rendering Web pages, including HTML5, CSS, SVG, Ajax, and other technologies for Web Applications;
- Web Architecture focuses on the foundation technologies and principles which sustain the Web, including URIs and HTTP;
- Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data are empowered by technologies such as RDF, SPARQL, OWL, and SKOS;
- XML Technologies including XML, XQuery, XML Schema, XSLT, XSL-FO, Efficient XML Interchange (EXI), and other related standards;
- Web of Services refers to message-based design frequently found on the Web and in enterprise software. The Web of Services is based on technologies such as HTTP, XML, SOAP, WSDL, SPARQL, and others;

– Web of devices. This includes Web access from mobile phones and other mobile devices as well as use of Web technology in consumer electronics, printers, interactive television, and even automobiles;
– Browsing and authoring tools. Web agents are intended to serve users.

5 Conclusion

The aim of the multimedia standards are to enable transparent and augmented use of multimedia resources across a wide range of networks, devices, user preferences, and communities, notably for trading of content.

References

1. Falchi, F., Gennaro, C., Savino, P., Stanchev, P.: Efficient Video Stream Filtering. IEEE Multimedia, 52–61 (January-March 2008)
2. Gelissen, J.: Introduction to MPEG-V. Journal of Virtual Worlds Research, Technology, Economy, and Standards 2(3), 4–7 (2009)
3. ISO/IEC 14496-11, Information technology, Coding of audio-visual objects, Part 11: Scene description and application engine, http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=54839
4. Janneck, J., Mattavelli, M., Raulet, M., Wipliez, M.: Reconfigurable Video Coding - a Stream Programming Approach to the Specification of New Video Coding Standards. In: MMSys 2010, February 22-23, pp. 223–234 (2010)
5. MHEG organization, http://www.mheg.org/users/mheg/index.php
6. MPEG organization, http://www.mpeg.org/
7. Ruhrmann, G., Nieland, J.: Interaktives Fernsehen. Westdeutscher Verlag GmbH, Wiesbaden (1997)
8. Stanchev, P., Amato, G., Falchi, F., Gennaro, C., Rabitti, F., Savino, P.: Selection of MPEG-7 Image Features for Improving Image Similarity Search on Specific Data Sets. In: 7th IASTED International Conference on Computer Graphics and Imaging, CGIM 2004, Kauai, Hawaii, pp. 395–400 (2004)
9. Stanchev, P.: ANIMATION, System for Animation Scene and Content Creation, Retrieving and Viewing. In: SPIE, vol. 4672, pp. 86–94 (2002)
10. Stanchev, P.: Semantic Video and Image Retrieval - State of the Art and Challenges. In: ACMSE 2007, The 45th ACM Southeast Conference, Winston-Salem, North Carolina, USA, March 24 (2007)
11. W3C organization, http://www.w3.org/standards/

Application of Wavelets and Kernel Methods to Detection and Extraction of Behaviours of Freshwater Mussels*

Piotr Przymus[1], Krzysztof Rykaczewski[1], and Ryszard Wiśniewski[2]

[1] Faculty of Mathematics and Computer Science,
Nicolaus Copernicus University, Toruń, Poland
[2] Laboratory of Applied Hydrobiology,
Nicolaus Copernicus University, Toruń, Poland

Abstract. Some species of mussels are well-known bioindicators and may be used to create a Biological Early Warning System. Such systems use long-term observations of mussels activity for monitoring purposes. Yet, many of these systems are based on statistical methods and do not use all the potential that stays behind the data derived from the observations. In the paper we propose an algorithm based on wavelets and kernel methods to detect behaviour events in the collected data. We present our algorithm together with a discussion on the influence of various parameters on the received results. The study describes obtaining and pre-processing raw data and a feature extraction algorithm. Other papers which applied mathematical apparatus to Biological Early Warning Systems used much simpler methods and their effectiveness was questionable. We verify the results using a system with prepared tags for specified events. This leads us to a classification of these events and creating a *Dreissena polymorpha* behaviour dictionary and a Biological Early Warning System. Results from preliminary experiments show, that such a formulation of the problem, allows extracting relevant information from a given signal and yields an effective solution of the considered problem.

Keywords: Automated biomonitoring, Biological Early Warning System, Wavelets, Time series, Zebra mussel (Dreissena polymorpha).

1 Introduction

Monitoring of water contamination is one of the most crucial aspects of environmental and health care. Many existing monitoring systems examine water only for a narrow range of substances and work without continuous control. For that reason, systems based on life organisms, i.e. *Biological Early Warning Systems* (BEWS), increasingly gain interest and popularity. Building BEWS is a complex

* This work was supported in part by the Marshall of Kuyavian-Pomeranian Voivodeship in Poland with the funds from European Social Fund (EFS) (a part of integrated operational program for regional development, activity 2.6) in the form of a grant for PhD students (Step in the future program, second edition).

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 43–54, 2011.

task, which requires a choice of a relevant bioindicator for the monitored environment, preparation of an activity measuring system, that will provide data for further processing, developing analysis and characterisation methods.

As aquatic organisms are sensitive to the concentration changes of different life supporting substances or the presence of xenobiotic, stressing or toxic compounds in the water, they are eligible as bioindicators. Most frequently used as sensing elements are cladocerans [1, 2], amphipods [3], bivalves [4, 5, 6], aquatic insects (*Chironomidae*) larvae [7] and fish [8]. Especially mussels, like *Mytilus* or *Dreissena*, as sessile bivalves, are very suitable for long-term, *in situ* water quality monitoring.

There are several methods for measuring the response of mussels to stressing factors. In older systems it was measured as a frequency of shell closing-opening events, through gluing wires to both halves of the shell and connecting them through the interface to a computer [9, 10]. The number of closed mussels in a treated group, in comparison to control, was a measure of stress response. More recently, a wire was replaced by a magnetic coil (or Hall sensor), on one valve, and a magnet on the other [11]. The value of the amplified signal was proportional to the distance (gape) between the two valves. As a response to stress of a tested group, the average value of gape in comparison to control mussels was measured. Both systems have limitations in informative and interpretative value of generated data. Our observations of *Dreissena polymorpha* mussels behaviour showed, that the response to stressing factor is more complex. The sequence of elementary events, i.e. an extent of gape change value and time of the return to the initial gape value can form specific patterns for various natural or stress caused activity rhythms. The presence of such rhythms was confirmed in [12].

In this paper, we propose an algorithmic, fully automated analysis method for extraction of the behaviour of the zebra mussels (*Dreissena polymorpha*). Because the behaviour of the zebra mussels is recorded as long series of shell states, logged every second, we needed an efficient analytic tools [7, 13]. For this reason, we applied mathematical apparatus of wavelets and kernel methods.

Detection of signal changes using the methods for spectral decomposition of time series is especially interesting. Fourier Transform (FT) technique can be applied to analyse the frequency spectrum, but it does not provide any insight into when a frequency component is present. In other words, we gain no information about either the time at which peak occurs or its duration in time (i.e. *localization in time*). Because of the limitations of the FT technique, we recommend using the wavelet transforms for investigating the long-term records of sudden changes in animal behaviour. Moreover, this approach may be used to dissecting the impact of unexpected events such as disruption in electric circuits.

The paper is organised as follows. In Section 2 we focus on obtaining and pre-processing raw data. Section 3 is devoted to give necessary background of wavelet theory. In Section 4 we present our behaviour extraction algorithm, which effectiveness is evaluated in Section 5. Finally, in two last Sections 6 and 7, we point out used programming tools and conclude discussing the results.

2 Materials and Methods

Biological Signal. Signal is a record of activity of freshwater mussels. We measure the changes of the distance between the valves. For a sample signal acquired from zebra mussel see Figure 2 on page 50.

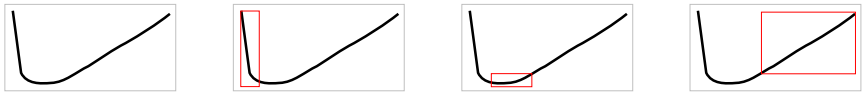
We want to extract single motions of mussels to classify them. It was proved in [12] that there are complex rhythms in the behaviour of *Dreissena polymorpha*. For example, the wanted pattern (shape of the graph of behaviour) may include the following phases: closing, resting and opening. These stages are presented in Figure 1 on page 46. Moreover, apart from the closing and opening phases, a vibration may occur. These are reactions to a stress or living activities. We search for time series fragments with following properties: at least closing and opening phases must appear, resting phase is optional; all phases may include perturbations. We analyse data with 16 minutes and 40 seconds (1000 seconds) periods of activity which from now on will be called *fragments*.

Measuring System. In our system, there are 8 mussels, which are located in a flow-through aquarium. They are attached to the ground, which does not affect their behaviour because of their sedentary nature.

We measure the changes of a magnetic field of a magnet placed on one side of the shell with a sensor placed on the other part of the shell. Data is collected every second from the sensors and transmitted to a database. The result sets showed, that the first prototype generates quite noisy data. Therefore, the measuring system should be improved in the future. The main difference between the old system and the new system will be based on used components type, their size and other resistance to interference from environment.

Obtaining and Pre-processing Raw Data. Denoising and data preparation steps consists of pre-processing filtering, removing white noise and averaging phase. A particularly important class of linear time-invariant systems are filters [14]. When the term frequency selective filter is used, it means that a system passes specific frequency components and totally rejects all others. This recommendation is common in particular for frequency selective filters like low-pass, band-pass and high-pass filters. A high-pass filter passes high frequencies well, but reduces the frequencies lower than the cut-off frequency. The real attenuation amount for each frequency differs from filter to filter. In our study, we used wavelet filter, which is high-pass filter and will be described later.

Analysis Method. Other papers investigated frequencies of closing-opening events [9, 10]. Previous results of observations conducted by the *Laboratory of Applied Hydrobiology at Nicolaus Copernicus University* reveal that one is able to extract motions as presented in Section 2 and, based on the activity record, is able to successfully assign water pollution to appropriate behaviour of mussel [9]. Stressful situation affects them, but it does not have to be a very violent reaction. For example, cyanotoxins and herbicides provide a recognisable, but not very intense, reaction. Therefore, we decided to analyse the normal activity



(a) Pattern (b) Opening (c) Resting (d) Closing

Fig. 1. Elementary phases of *Dreissena polymorpha* behaviour from time series point of view

and activity in a stressful situation to determine whether there is a difference and how it emerges.

One approach is to use statistical deviation. As abnormal we can take something that happens infrequently [8]. Not all incidents of behaviour, that do not conform to the norm, show an abnormality that we are looking for. We try to identify the *frame* (which will be defined in Section 4) in terms of shape, but not in terms of intensity or range of temporary phenomena.

In the future, we want to go even further and beyond the aforementioned situations to take an analysis of the captured events.

3 An Overview of the Wavelet Theory

The fundamental theory of wavelets was put forward by Haar in 1909 and then developed at the end of the 1960s. Now it has been extensively documented [15]. Wavelets have been very successful as an analytical tool to represent signals, in denoising, data compression and in time-scale analysis of time series, to mention a few of their applications. We refer inquisitive reader for more details concerning wavelet theory to [16].

Results obtained by this method are better than these obtained by Fourier analysis or other filter methods [15]. Because wavelet transforms can be exploited to analyse even non-stationary signals, they have been used in many medical applications and have been successful alternative methods to Fourier analysis.

Wavelets, in contrast to the Fourier Transform, are examples of *Multiscale Resolution Analysis*, which means that wavelet coefficients contain at once information about the frequency and the time domain. Thanks to this, they are particularly useful where the knowledge of these two characteristics of the signal is needed at the same time [16].

Continuous Case. Historically, wavelet analysis begins with *Continuous Wavelet* Transform (CWT). It provides a time-scale representation of a continuous function where the scale plays a role analogous to the one of the frequency in the analysis with the well-known Fourier Transform (FT). The main wavelet (the so-called *mother wavelet*) is a real valued function, that satisfies the following relations: $\int |\psi(t)|^2 dt = 1$ (quickly disappears), $\int \psi(t) dt = 0$ (oscillates). There are two main operations on wavelets: shift and rescaling. By applying them to

the mother wavelet we obtain a whole family of wavelets: for $j, k \in \mathbb{Z}$ and a wavelet ψ, let $\psi_{j,k}$ stands for the wavelet with scale j and displacement k, i.e.

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j \cdot t + k). \tag{1}$$

Wavelet transform is a mapping which assigns to a 1-dimensional signal $f(t)$ a 2-dimensional array $c_{j,k}$ in the following way

$$f(t) = \sum_{j,k} c_{j,k}\psi_{j,k}(t). \tag{2}$$

At every step of analysis, we have a convolution (which is a filter) and rescaling (n and t become $2n$ and $2t$, respectively).

By introducing the so-called *scaling function* ϕ_k (for more details see [16]) one can represent a signal as

$$f(t) = \sum_k b_k\phi_k(t) + \sum_j \sum_k d_{j,k}\psi_{j,k}(t) \tag{3}$$

where $d_{j,k} = \langle g, \psi_{j,k}\rangle = \int f(t)\overline{\psi_{j,k}(t)}\,dt$. The first sum represents the *approximation* A_j of a signal f at the level j which is given by the scaling function. The second sum represents the *detail* D_j at the level j and is given by wavelets. The key idea of the multiresolution is a decomposition of the signal into different scales and its reconstruction from the sum [15], e.g. $D_1+D_2+D_3+D_4+D_5+A_5$.

Discrete Wavelet Transform. Discrete Wavelet Transform (DWT) is a discrete version of the CWT, analogously like Discrete Fourier Transform (DFT) is a discrete version of the FT. In the equation (2) the DWT is given by the set of coefficients $c_{j,k}$.

The basic tool of wavelet analysis is a multiscale decomposition of the signals, which is implemented using multi-band wavelet complementary filters (high-pass filters and low-pass filters). Calculation procedure leading to this decomposition is called *Mallat algorithm* [15]. This algorithm allows a fast wavelet decomposition and a reconstruction of a signal.

Wavelet decomposition may be seen as a continuous time wavelet decomposition sampled at different frequencies at every level of the analysis. A suitable way to find the best level for the hierarchy, depends on the signal type. In general, the level is selected depending on the desired low-pass cut-off frequency [7].

Analysis Using Wavelet Packet Transform: Tuning of the Various Levels. Discrete Wavelet Transform (DWT) is a particular case of *wavelet packet analysis* [16]. Moreover, implementation of wavelet packet analysis is done by dividing whole time-frequency into smaller pieces. The main reason for taking wavelet packet (WP) into consideration is to be able to analyse non-stationary signals and their behaviour.

Selection of Appropriate Wavelets for the Considered Problem. Several different families of wavelet functions have been defined [15]. Each of them is characterised by different properties, such as smoothness, compact support, and so on. In our case, the selection of appropriate wavelet is done by the algorithm.

4 Event Extraction Algorithm

Let us now present our algorithm, which captures the events. Then we justify its correctness. Having a filtered signal we are trying to cut it into elementary events and analyse them. The algorithm was created in a parametric form. There are following parameters: `name`, `level`, `local_error_frame_size`, `box_cleanup`, `box_threshold`. Below the meaning of these parameters is discussed.

Notation. We analyse signal which is assumed to be a time series $\{x_i = x(t_i)\}_{i\in J} \subset \mathbb{R}$, where $T = \{t_i\}_{i\in J}$ is a discrete set of times and $J \subset \mathbb{N}$ is finite set of indices.

Each subset $F \subset T$ having the property

$$\text{if} \quad t_i \in F, t_k \in F \text{ with } i < k, \quad \text{then} \quad \forall_{i<j<k}\, t_j \in F \tag{4}$$

is called *event*.

Given $F = \{t_k\}_{k=k_0}^{k_1}$ by *frame of an event* we understand the set

$$F \times \left[\min_{k\in\{k_0,\ldots,k_1\}} \{x(t_k)\}, \max_{k\in\{k_0,\ldots,k_1\}} \{x(t_k)\}\right]. \tag{5}$$

Further, by *behaviour extraction* we understand extraction of events with the desired properties as presented in Section 2.

Construction of the Filter. Firstly, we prepare the data: we unfold the data and remove noise by using a *wavelet filter*. A wavelet filter is a non-linear digital filtering technique, usually necessary to perform a high degree of noise reduction in a signal, before performing higher-level processing steps. This filter turns off a signal component at a certain level of wavelet analysis, i.e. it sets out $D_n = 0$ in the reconstruction step. In our case filter have two parameters: `filter_name` (specifies the name of used wavelet in this filter) and `filter_level` (specifies which component of the signal is turned off).

Local Error Estimator. In Figure 2, in addition to the high frequency components of the signal, we show a plot of a function, which is proportional to the absolute value of *the kernel weighted average* (*the Nadaraya-Watson kernel regression estimator*), in a neighbourhood of each point x_i. It is the convolution of Gaussian density function $\eta(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2}$ and the signal x, which in the discrete case is given by

$$k_x(x_{i_0}) = [\eta * x](x_{i_0}) = \sum_{i\in J} \frac{e^{-\frac{(i_0-i)^2}{\sigma}}}{\sqrt{2\pi}} \cdot x_i, \tag{6}$$

where $\sigma =$`local_error_frame_size`, $x_{i_0} \in x(T) := \{x_i\}_{i\in J}$. For more details and other kernels see [17]. At this stage, one could also calculate a common mean, i.e. $l(x_{i_0}) = \frac{1}{|J|}\sum_{i\in J} |x_{i_0} - x_i|$. This, however, does not take into account the local behaviour and gives, therefore, worse results.

Main Idea behind the Algorithm. The algorithm, which will be presented below, enables us to detect sudden jumps and sharp cusps in a time series by using a discrete wavelet transform. The idea is simple: a sudden jump of the time series affects the magnitudes of wavelet coefficients, so one can set a threshold level to find the point at which the jump occurs.

After decomposing a signal by using the wavelet packet with a wavelet function given by the parameter `name`, we search for interesting events. This computation can be described as follows. We consider detail of the given signal, which will be denoted by D, at the level which is given by the parameter `level`. According to our observations, in the component D there are a lot of information about the signal (sudden jumps, vibrations, fluctuations). Let us define x_+, x_- as positive and negative part of considered component, i.e.

$$x_+ = \big\{ \max(v, 0) \mid v \in D \big\}, x_- = \big\{ \max(-v, 0) \mid v \in D \big\}. \tag{7}$$

Plots of x_- and x_+ are shown in Figure 2 as continuous lines on the subfigures `start` and `end`. Let us define the *START* and *END* flags:

$$\begin{aligned} \mathsf{START} &= \big\{ i \in J \mid (x_-[i-1] < k_{x_-}(x_i) \leqslant x_-[i]) \wedge (k_{x_-}(x_i) > w) \big\}, \\ \mathsf{END} &= \big\{ i \in J \mid (x_+[i-1] > k_{x_+}(x_i) \geqslant x_+[i]) \wedge (k_{x_-}(x_i) < w) \big\}, \end{aligned}$$

where k_{x_-} and k_{x_+} (dotted curves in Figure 2 on subfigures `start` and `end`) are kernel weighted averages for x_- and x_+ respectively (see 6) and $w =$ `box_threshold`.

The analysed component D may have a big disruption and this may result in frequent occurrence of events. It can be seen, that the parameter w provides a barrier beyond which the events occurred. Moreover, it prevents too frequent appearance of events. Here we find places where the signal is above or below the kernel weighted average at a given point, which is defined by formula (6). These points are suspected of being starts and ends of the frames.

Finding the best coverage by the frames using START and END flags can be done in the following way:

$$\begin{aligned} \mathcal{S} &= \big\{ i \in \mathsf{START} \mid \exists_{k \in \mathsf{END}} \neg\exists_{j \in \mathsf{START}}\, k < j < i \big\}, \\ \mathcal{E} &= \big\{ i \in \mathsf{END} \mid \exists_{k \in \mathsf{START}} \neg\exists_{j \in \mathsf{END}}\, k > j > i \big\}. \end{aligned}$$

The first element of START is supposed to be in the set $\mathcal{S}$. Further, one can show that the sets $\mathcal{S}$ and $\mathcal{E}$ sets contain the same number of elements. The sets $\mathcal{S}$ and $\mathcal{E}$ are declared to be the points at which opening and closing phases occurs, respectively.

In Figure 2, we can also see dotted vertical lines which represent points, that were suspected to be in the classes $\mathcal{S}$ and $\mathcal{E}$, but were omitted by this algorithm. Analysis of Figure 2 justifies the choice of these sets. Thus, we obtain the events

$$\big[\mathcal{S}(i_1), \mathcal{E}(i_1)\big], \big[\mathcal{S}(i_2), \mathcal{E}(i_2)\big], \ldots, \big[\mathcal{S}(i_q), \mathcal{E}(i_q)\big] \tag{8}$$

that have to be compared (which in general vary in length). Now we have to check if the selected events indeed generate good frames, i.e. if the height of a frame $>$ `box_cleanup` $\times$ width of a frame. This algorithm is greedy. Therefore, we introduce parameter `box_cleanup`, which protects from taking into one event the whole fragment. Figure 2, shows the detected frames (last graph in each picture).

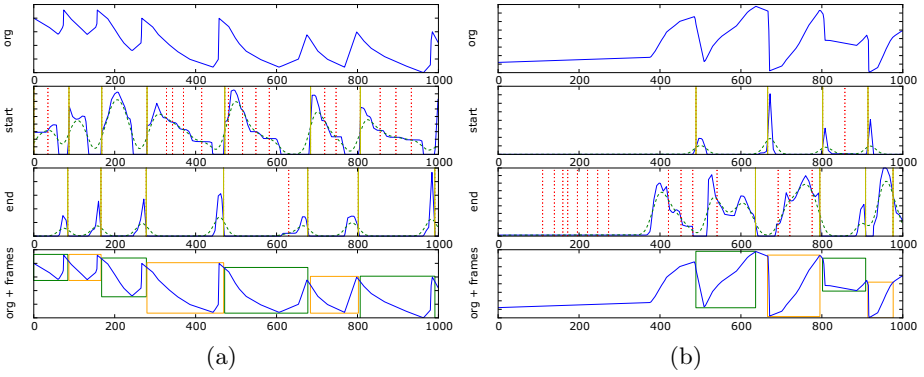


Fig. 2. Frames extracted using our algorithm

5 Experiment Results

In this section, we will evaluate the effectiveness of the proposed algorithm, against the real data. All data were derived from experiments, in which impact of salt, herbicide and yeast on mussels was tested.

Data Specification. We used the data from 8 sensors with frequency of reading being 1 second. Each sensor recorded 11245 minutes of data which consist of reals from the interval $[0, 45]$. Sensor read error is assumed to be in the interval $[0, 3]$. Due to technical requirements the data set was split into time windows (fragments) of length 1000 seconds.

It is important to note that this data set is a subset from a bigger set of experiments, i.e. besides limiting data set, we only choose time fragments where height of the frame $>$ sensor read error.

Thus, in what follows, it is assumed, that everything what is below sensor read error is only a negligible perturbation. Considering this, we suggest using a more accurate sensor system. Some early experiments with better sensors (Hall sensors) show that our algorithm gains effectiveness.

The mussels were derived from the same colony. All biological experiments were conducted with at least 50% of mussels in control terms. There were also biological experiments conducted only in control conditions. Table 1 describes data set, obtained from this experiments.

Result Evaluation. The data set was independently analysed by two researchers who put markers in the areas of events described above. For comparing similarity between researchers and the algorithm we use *Tversky index* defined as follows:

$$S(X, Y) = \frac{\sum_{i,j} |X_i \cap Y_j|}{\sum_{i,j} (|X_i \cap Y_j| + \alpha |X_i - Y_j| + \beta |Y_i - X_j|)},$$

Table 1. Sum of read time for all mussels fulfilling conditions

Stressor	Herbicide	Yeast	Salt	Control
no oxygen	0	0	0	853 min
normal	853 min.	357 min.	986 min.	8193 min.

where $\alpha+\beta = 1$, X_i is i-th event and $|X_i|$ is its length. This asymmetric similarity measure, compares a *variant* to a *prototype*. We use values $\alpha = 0.75$ and $\beta = 0.25$.

With an algorithm constructed in such a way, we find the optimal parameter values using 30% of the data and event markers.

We measure similarity between markers provided by researchers and algorithm results. Moreover, we consider markers, of the researches, as a prototype and compare it using Tversky index with the algorithm results. Additionally, as Tversky index is asymmetrical, we also considered the case when algorithm is a prototype and researchers markers are a variant. See results in Table 2.

As a side note, it seems worth mentioning that the presented method is quite fast. For example, the data, obtained from 7375 minutes of the observations, is calculated in approximately 38.5 seconds (results were obtained on an Intel® Core™ 2 Quad CPU Q6600 2.40GHz).

Correctness of the Results. When choosing the best wavelet, which is then applied to the analysis of specified signal, shown in Figure 2, one should be guided by the well-known rule (see [15]), that the wavelets of "smooth" shape (e.g. `Morlet wavelet`) are characterised by better resolution in analysing signals in terms of spectrum, i.e. they are characterised by a better localization of frequency components on the frequency axis. Wavelets which are discontinuous or have slopes (e.g. `Haar wavelet`, `biorthogonal wavelet`), show better localization on the timeline. Therefore, the intuitions suggest that wavelets that are suitable for our purposes should have sharp cusps, which will caught a sudden jump in the signal. Indeed, at the stage of the automatic selection of the parameters, we obtained confirmation of our predictions.

Figure 2 clearly shows the characteristic signal in different frequency ranges, the amplitude and the location on the timeline. A notable feature is that the components with higher frequencies are concentrated in the relatively short length of time. The location agrees with the times of initiation or declines of movement of the mussel. This feature, which manifested distinct "peaks" in moments, corresponding to a behaviour phenomena are particularly evident at different levels of details. These facts give rise to the choice of beginnings and ends of single events in our system and explain our algorithm.

Optimization Results. During the optimization we obtained, the following parameters: `filter_name` = `db1`, `filter_level` = `aaad`, `name` = `rbior3.1`, `level` = `ddda`, `local_error_frame_size` = 450, `box_cleanup` = 0.12, `box_threshold` = 0.001. We see that the best results were obtained for the `rbio` wavelets, which have the above-mentioned properties (see [16]).

Table 2. Similarity between researches and algorithm clustered into cases of stressors

Experiment Description	User	User 1	User 2	Algorithm
Summary	User 1	x	0.712046377833	0.761884176385
	User 2	0.808793419224	x	0.715760702623
	Algorithm	0.648714212421	0.531441669327	x
Herbicide	User 1	x	0.73704199884	0.702517899468
	User 2	0.807069914053	x	0.716061338472
	Algorithm	0.666851522047	607025636399	x
Salt	User 1	x	0.685833427745	0.71389773682
	User 2	0.763362050728	x	0.629846521212
	Algorithm	0.677996432575	0.566973527048	x
Yeast	User 1	x	0.599296680011	0.647952077503
	User 2	0.766786953972	x	0.785170581729
	Algorithm	0.651532190762	0.623698430374	x
Control	User 1	x	0.712046377833	0.745311306628
	User 2	0.808793419224	x	0.766048200154
	Algorithm	0.705921742781	0.627110395079	x
Herbicide, no oxygen	User 1	x	0.73704199884	0.702517899468
	User 2	0.807069914053	x	0.716061338472
	Algorithm	0.666851522047	0.607025636399	x

6 Software Choices

Prototype implementation has been prepared in Python using `SciPy`, `NumPy`, `PyWavelets`, `MLPY`, `Matplotlib`, `flot` and `django`. `SciPy` is a Python library for mathematics, science and engineering. `NumPy` provides a library for convenient and fast N-dimensional array manipulation. Additionally, we used `PyWavelets`, a Discrete Wavelet Transform library to Python and `MLPY` – a machine learning library based on `NumPy` and GNU Scientific Library. The plots were prepared using `Matplotlib` and `flot`, the management layer and the experimental platform were prepared in `django`.

7 Conclusions and Future Work

Stressful situations can change the behaviour of mussels (both at the level of fundamental changes in the behaviour — anomalies can occur — and a rhythm disturbance of these behaviours). It can also cause an emergence of a new behaviour (e.g. testing the surrounding environment). The preliminary observations suggest, that there is a possibility of creating an alphabet of normal and abnormal behaviours, which will have to be extended depending on the stressful situations. Clustering of frames is possible and reasonable, but still some analysis details have to be modified to guarantee good results. It may be a good starting point for classification procedures, which will work very effectively.

Contributions of This Work

- We have developed a fast algorithm based on wavelets and kernel methods for the extraction of behaviours which in the future are going to be classified depending on the stressful situations.
- We have evaluated the effectiveness of the algorithm.
- We have developed a platform for an automatic behaviour detection and extraction.

Future Work. Proper functioning of this system requires gathering large quantities of mussels activities in natural conditions under high-stress factors and stressful conditions. Our further work will concentrate on the improvement of the clustering process, building of the alphabet and classifying the behaviours. We plan a set of experiments in laboratory and natural environment.

References

[1] Hendriks, A.J., Stouten, M.D.A.: Monitoring the response of microcontaminants by dynamic daphnia magna and leuciscus idus assays in the rhine delta: biological early warning as a useful supplement. Ecotoxicol. Environ. Safety 26, 265–279 (1993)

[2] van Hoof, F., Sluyts, H., Paulussen, J., Berckmans, D., Bloemen, H.: Evaluation of a bio-monitor based on the phototactic behavior of daphnia magna using infrared detection and digital image processing. Water Sci. Technol. 30, 79–86 (1994)

[3] Gerhardt, A., Carlsson, A., Resseman, C., Stich, K.-P.: New online biomonitoring system for gammarus pulex (crustacea): in situ test below a copper effluent in south sweden. Environ. Sci. Technol. 32, 150–156 (1998)

[4] Sloof, W., de Zwart, D., Marquenie, J.M.: Detection limits of a biological monitoring system for chemical water pollution based on mussel activity. Bull. Environ. Contam. Toxicol. 30, 400–405 (1983)

[5] Sluyts, H., van Hoof, F., Cornet, A., Paulussen, J.: A dynamic new alarm system for use in biological early warning systems. Environ. Toxicol. Chem. 15, 1317–1323 (1996)

[6] Borcherding, J., Jantz, B.: Valve movement response of the mussel dreissena polymorpha - the influence of ph and turbidity on the acute toxicity of pentachlorophenol under laboratory and field conditions. Ecotoxicology 6, 153–165 (1997)

[7] Kim, C.-K., Kwak, I.-S., Cha, E.-Y., Chon, T.-S.: Implementation of wavelets and artificial neural networks to detection of toxic response behavior of chironomids (chironomidae: Diptera) for water quality monitoring. Ecol. Model. 195, 61–71 (2006)

[8] Kramer, K.J.M., Botterweg, J.: Aquatic biological early warning systems: an overview. In: Jeffrey, D.J., Madden, B. (eds.) Bioindicators and Environmental Management, pp. 95–126. Academic Press, London (1991)

[9] Wiśniewski, R.: New methods for recording activity pattern of bivalves: A preliminary report on dreissena polymorpha pallas during ecological stress. In: Tenth Intern. Malacol. Congress, pp. 363–365 (1991)

[10] Borcherding, J.: Ten years of practical experience with the dreissena-monitor, a biological early warning system for continuous water quality monitoring. Hydrobiologia 556, 417–426 (2006)

[11] Pynnönen, K.S., Huebner, J.: Effects of episodic low ph exposure on the valve movements of the freshwater bivalve anodonta cygnea. L. Wat. Res. 29, 2579–2582 (1995)
[12] Gudimov, A.V.: Elementary behavioral acts of valve movements in mussels (mytilus edulis l.). Doklady Biological Sciences 391, 346–348 (2003); Translated from Doklady Akademii Nauk 391(3), 422–425 (2003)
[13] Rodland, D.L., Schöne, B.R., Helama, S.O., Nielsen, J.K., Baier, S.M.: A clockwork mollusc: Ultradian rhythms in bivalve activity revealed by digital photography. J. Exp. Biol. Ecol. 334, 316–323 (2006)
[14] Rutkowski, L.: Adaptive Filters and Adaptive Signal Processing (in Polish). WNT, Warsaw (1994)
[15] Chui, C.K.: Wavelets: A mathematical tool for signal analysis. SIAM, Philadelphia (1997)
[16] Wojtaszczyk, P.: A Mathematical Introduction to Wavelets. Cambridge University Press, Cambridge (1997)
[17] Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, Oxford (1995)

Test Cost Constraint Reduction with Common Cost

Guiying Pan, Fan Min*, and William Zhu

Lab of Granular Computing, Zhangzhou Normal University, Zhangzhou 363000, China
minfanphd@163.com

Abstract. Test cost is an important issue in cost-sensitive systems. It is what we pay for obtaining a data item of an object. In some applications, there are some common costs and a cost constraint. The common cost is due to the share of the same resources by several tests. The cost constraint is due to limited money, time, or other resources. Recently, the two issues have been addressed independently in cost-sensitive rough sets. In contrast, this paper considers both issues. Our problem is to construct test sets meeting the constraint and preserving the information of decision systems to the highest degree. We propose a heuristic algorithm to deal with this problem. It is based on information gain, test costs, group-memberships, common costs and a non-positive exponent λ. λ is employed in the penalty function such that expensive tests are unlikely to be chosen. Experimental results indicate that the algorithm performs good in terms of the possibility of finding the optimal reduct. Since the optimal setting of λ is often unknown, we can run the algorithm with different λ values and obtain better results.

Keywords: Cost-sensitive learning, test cost, common cost, constraint, heuristic algorithm.

1 Introduction

Cost-sensitive learning is one of the most challenging problems in data mining applications [18]. The research work on cost-sensitive learning is fruitful in different data mining domains, such as rough sets [6,7,20,21], decision trees [16], artificial neural networks [5,23], and Bayes networks [1].

In test-cost-sensitive decision systems, attributes of an object are called tests. Test cost is what we pay for collecting a data item. It can be money, time, or other types of resources. In many applications, such as clinical systems, to save resources, we wish to undertake only a part of tests. This issue has been formalized as the minimal test cost reduct (MTR) problem in [6]. However, due to limited money, time, or other resources, we sometimes have a constraint on the total test cost. We have to select a test set which meets the constraint and meanwhile preserves the information of decision systems to the highest degree. This problem is referred as the search of the optimal sub-reducts with test cost constraint (OSRT) problem [11].

There is often a certain associative relation among some tests. For example, the cost of collecting a blood sample from a patient is shared by the set of blood tests [7]. We say

* Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 55–63, 2011.

these tests share a common cost. The simple common test cost decision system (SCTC-DS) and the corresponding attribute reduction algorithm were introduced in [2]. The SCTC-DS is more general than the test-cost-independent decision system (TCI-DS) [6]. Thus, the OSRT problem is more challenging and interesting in case of the SCTC-DS.

In this paper, we propose a heuristic algorithm to deal with the new problem. Due to the complexity of the SCTC-DS, the approach to compute total test cost is different from one in the TCI-DS. Besides test costs, we need to consider group-memberships and common costs. We use information gain as the heuristic information and meanwhile consider test costs, group-memberships and common costs. We set a user-specified non-positive exponent λ to adjust the influence of test costs. When $\lambda < 0$, expensive tests are more unlikely to be chosen.

The heuristic algorithm is tested for the Mushroom dataset with three different data distributions [6], namely Uniform, Normal, and Bounded Pareto distributions. We study how frequent the algorithm can find the optimal solution. This evaluation measure is referred to as the finding optimal factor (FOF) [6]. The algorithm is implemented in our open source software Coser [12]. Experimental results show that the heuristic algorithm produces good results especially for the Uniform distribution. Since the optimal setting of λ is often unknown, we can run the algorithm with different λ values and obtain better results. We observe that if λ influences the quality of the result significantly, then the approach is more likely to help increasing the performance of the algorithm.

2 Preliminaries

This section reviews some preliminary knowledge of the paper. First, the data model is described; it is represented a decision system, a cost vector, a group-membership function and a group common cost function. Then the OSRT problem is introduced.

2.1 Simple Common Test Cost Decision System

Most data mining and machine learning are based on the decision systems. The decision system is often defined as follows.

Definition 1. *[19] A decision system (DS) S is the 5-tuple:*

$$S = (U, C, D, V, I), \tag{1}$$

where U is a finite set of objects called the universe, C is the set of conditional attributes, D is the set of decision attributes, $V = \{V_a | a \in C \cup D\}$ where V_a is the set of values for each $a \in C \cup D$, and $I = \{I_a | a \in C \cup D\}$ where $I_a : U \to V_a$ is an information function for each $a \in C \cup D$.

A decision system is generally represented by a decision table as Table 1.

Test-cost-independent decision systems (TCI-DS) are simpler and more widely used models extended from decision systems. Let $S = (U, C, D, V, I, c)$ be a TCI-DS, where U, C, D, V and I are the same meaning as in Definition 1, c is an attribute cost function represented by a vector $c = [c(a_1), c(a_2), \ldots, c(a_{|C|})]$. For example,

Table 1. An exemplary decision table

Patient	Headache	Temperature	Lymphocyte	Leukocyte	Eosinophil	Heartbeat	Flu
x_1	yes	high	high	high	high	normal	yes
x_2	yes	high	normal	high	high	abnormal	yes
x_3	yes	high	high	high	normal	abnormal	no
x_4	no	high	normal	normal	high	normal	no

Table 2. An exemplary group-membership vector

a	Headache	Temperature	Lymphocyte	Leukocyte	Eosinophil	Heartbeat
$g(a)$	1	1	2	2	2	3

$c = [12, 15, 15, 20, 15, 10]$ indicates that the test cost of Headache, Temperature, Lymphocyte, Leukocyte, Eosinophil, Heartbeat are \$1, \$1, \$2, \$2, \$2, and \$3, respectively. For any $B \subseteq C, c(B) = \sum_{a \in B} c(a)$. A simple common test cost decision system is more general than a test-cost-independent decision system, defined as follows.

Definition 2. *[7] A simple common test cost decision system (SCTC-DS) S is the 8-tuple:*

$$S = (U, C, D, V, I, c, g, gc), \tag{2}$$

where U, C, D, V, I, and c have the same meaning as in the definition of the TCI-DS, $g : C \to [1, \ldots, K], 1 \leq K \leq |C|$, is the simple group-membership function, and $gc : [1, \ldots, K] \to \mathbb{R}$ is the group common cost function. The k-th group $(1 \leq k \leq K)$ is given by $G_k = \{a | g(a) = k\}$, and attributes in the group share a common cost $0 < gc(k) \leq min_{a \in G_k} c(a)$.

The group-membership function and the group common cost function can be represented by vectors $g = [g(a_1), g(a_2), \ldots, g(a_{|C|})]$ and $gc = [gc(1), gc(2), \ldots, gc(|K|)]$. An example group-membership function is listed in Table 2, and a group common cost function is given by $gc = [2, 2, 3]$. That is, the common costs for three groups are \$2, \$2, and \$3, respectively. When $|K| = |C|$, the SCTC-DS coincides with the TCI-DS. For any $B \subseteq C$ and $a \in G_k - B$,

$$c^*(B \cup \{a\}) = \begin{cases} c^*(B) + c(a) - gc(k), & \text{if } B \cap G \neq \emptyset; \\ c^*(B) + c(a), & \text{otherwise.} \end{cases} \tag{3}$$

2.2 Optimal Sub-reducts with Test Cost Constraint Problem

Attribute reduction is to remove redundant attributes and keep the ability of decision systems as original decision systems. Many data models such as classical [13], covering-based [24,25], variable-precision [26], and neighborhood [3] rough sets are proposed to address this problem. A number of approaches to relative reduction are proposed,

such as positive region reduction [13,14], maximum distribution reduction [22], Shannon's entropy reduction [17], fuzzy reduction [4] and β-reduction [26] and so on. In this paper, we employ Shannon's entropy reduction which is represented as follows. Let S be a decision system, and $H(D|B)$ be the conditional entropy of $B \subseteq C$ w.r.t. D. Any $B \subseteq C$ is a Shannon's entropy reduct if and only if $H(D|B) = H(D|C)$ and $\forall a \in B, H(D|B - \{a\}) > H(D|B)$. Any $B \subseteq C$ is called a sub-reduct if $\forall a \in B, H(D|B - \{a\}) > H(D|B)$ [10].

In decision systems, relative reduction is to find the minimal description of the data. Generally, the objective is to find attribute subsets with the minimal size [15]. However, in test-cost-sensitive decision systems, the minimal reducts are not always the optimal reducts with minimal test cost. Thus, the minimal test cost attribute reduct (MTR) problem is proposed in [6]. Let $Red(S)$ denote the set of all relative reducts of a test-cost-sensitive decision system S. Any $R \in Red(S)$ where $c(R) = min\{c(R')|R' \in Red(S)\}$ is called a minimal test cost reduct.

In some complicated situations, due to limited money, time, or other resources, we have to select a test subset. This subset meets the constraint and meanwhile preserves the information of the decision system to the highest degree. This problem is called optimal sub-reducts with test cost constraint (OSRT) problem which is defined as follows.

Definition 3. *[10] Let $S = (U, C, D, V, I, c)$ be a TCI-DS and m be the test cost upper bound. The set of all test sets subject to the constraint is*

$$T(S, m) = \{B \subseteq C | c(B) \leq m\}. \tag{4}$$

In $T(S, m)$, the set of all test sets with the minimal conditional entropy is

$$M_T(S, m) = \{B \in T(S, m) | H(D|B)\} = min\{H(D|B') | B' \in T(S, m)\}. \tag{5}$$

In $M_T(S, m)$, the set of all optimal sub-reducts is

$$P_{M_T}(S, m) = \{B \in M_T(S, m) | c(B)\} = min\{c(B') | B' \in M_T(S, m)\}. \tag{6}$$

Any element in $P_{M_T}(S, m)$ is called an optimal sub-reduct with test cost constraint, or an optimal sub-reduct for brevity.

It is obvious that the OSRT problem is more general than the MTR problem.

3 The Algorithm

To evaluate the performance of the heuristic algorithm, the optimal solution should be obtained first. Similar to BASS [9], we designed a backtrack algorithm and implemented it in Coser [12].

The focus of the paper is a heuristic algorithm as listed in Algorithm 1. We consider information gain, test costs, group-memberships and common costs. We also use the competition approach [6,8] to select the best sub-reduct. Note that the best one produced by the algorithm may not be an optimal one.

Let $B \subset C$ and $a_i \in C - B$, the information gain of a_i w.r.t. B is

$$f_e(B, a_i) = H(D|B) - H(D|B \cup \{a_i\}). \tag{7}$$

Let a_i be a member of the k-th group G_k, $gc(k)$ be the common cost of G_k and λ be a non-positive value, the λ-weighted function is defined as

$$f(B, a_i, c) = \begin{cases} f_e(B, a_i)c_i^{\lambda}, & G_k = \emptyset; \\ f_e(B, a_i)(c_i - gc(k))^{\lambda}, & G_k \neq \emptyset, c_i \neq gc(k). \end{cases} \tag{8}$$

Note that if $G_k \neq \emptyset, c_i = gc(k)$, a_i is selected without any cost at once.

4 Experiments

The experiment is tested 100 times on the Mushroom dataset with different setting. Three data distributions, namely Uniform, Normal and Bounded Pareto distribution, are employed to produce test costs. They have the same settings as those in [6]. Test costs are random integers ranging from 1 to 100. The α is 8 and 2 for the Normal distribution and the Bounded Pareto distribution, respectively. Group-memberships and common costs meeting Definition 2 are also randomly produced. The group-memberships are set simply here, denoted by the number of group (NOG). The NOG ranges from 1 to $|C|$ which is 22 in Mushroom dataset. Therefore each group has $|C|/NOG$ tests on average. $m = \lfloor c(R^*) \times 0.8 \rfloor$ where R^* is the minimal test cost reduct, and $\lambda \in L$ which ranges from 0 to -2 with a step length 0.25.

Let the number of experiments be K and the number of successfully finding an optimal solution be k, the finding optimal factor (FOF) which is defined as k/K [6]. Table 3 shows partial results of influence of group-memberships and different λ on the FOF. Fig 1(a) depicts the final results from global perspective. Figure 1(b), 1(c) and 1(d) depict the influence of λ on the results from the local perspective. We have the following observations.

1. λ influences the quality of the result significantly, and the final result produced the competition is generally good. When the NOG is fixed, the results produced with different λ are most different. That is, the color at the same NOG is changeful in the three figures. Then the competition approach can improve the final result greatly.
2. As a whole, the final result in the Uniform distribution is more stable than that in other two distributions. The maximal and minimal value of the FOF in the Uniform are 0.85 and 0.75, respectively. The reason is that λ makes bigger influence on results in the Uniform distribution than in others on the whole.
3. The curve variations of the FOFs in the Normal and Bounded Pareto are similar with NOG from 5 to 22 expect 12. Most test costs distribute from 40 to 60 in the Normal distribution and their mean is nearby 50. Most test costs are 1 in the Bounded Pareto distribution and their mean is nearby 2. These two distributions have a similarity that their data are relative centralized to respective mean so that they have a similar curve variation. However, since the test costs produced by the former are generally larger than ones produced by the latter. The influence of λ in the former is larger than that in the latter. Therefore, the final result in the Normal distribution is better than that in the Bounded Pateto distribution.

Algorithm 1. A heuristic algorithm for OSRT problem in SCTC-DS

Input: $S = \{U, C, D, V, I, c, g, gc\}$, m
Output: A suboptimal sub-reduct with test cost constraint
Method: SC-OSRT reduction

```
1: B = ∅;//The sub-reduct.
2: CA = C;//Unprocessed attributes.
3: G_k = ∅;//Processed attributes in the k-th group.
4: tc = m;//Available test cost upper bound.
5: while (CA ≠ ∅) do
6:    for (each a ∈ CA) do
7:       k = g(a);c'_a = c_a;
8:       if (G_k ≠ ∅) then
9:          c'_a = c_a − gc(k);
10:      end if
11:      if (c'_a = 0) then
12:         B = B ∪ {a}; CA = CA − {a}; continue;
13:      end if
14:      if (c'_a ≤ tc) then
15:         compute f(B, a, c'_a);
16:      else
17:         CA = CA − {a};
18:      end if
19:   end for
      //Addition
20:   Select a' with the maximal f(B, a', c'_a');
21:   B = B ∪ {a'}; CA = CA − {a'}; k = g(a'); G_k = G_k ∪ {a'};
      //Deletion, B must be a sub-reduct.
22:   CD = B;Sort attributes in CD according to respective test cost in a descending order;
23:   while (CD ≠ ∅) do
24:      CD = CD − {a'};//a' is the first element of the set CD.
25:      if (H(D|B − {a'}) = H(D|B)) then
26:         B = B − {a'}; k = g(a'); G_k = G_k − {a'};
27:      end if
28:   end while
29:   Compute c_B;
30:   tc = m − c_B;
      //Is B a reduct?
31:   if (H(D|B) = H(D|C)) then
32:      break;
33:   end if
34: end while
35: return B;
```

4. When the NOG is 1 or 2, the FOF in the Bounded Pareto distribution is lower than 0.15. With this distribution, the test cost of optimal reducts is often 1 so that the

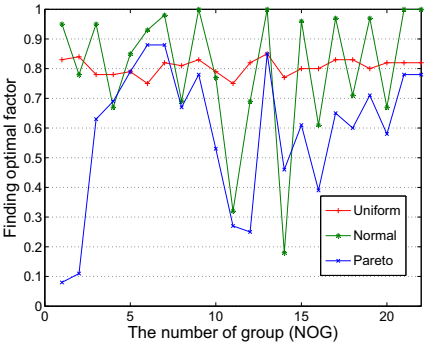
Table 3. The FOF based on group-memberships and λ on Mushroom dataset

Distribution	λ	Groups										
		2	4	6	8	10	12	14	16	18	20	22
Uniform	0	0.50	0.45	0.42	0.55	0.48	0.46	0.51	0.49	0.55	0.49	0.48
	-0.25	0.60	0.48	0.52	0.60	0.56	0.52	0.54	0.53	0.62	0.53	0.52
	-0.5	0.60	0.56	0.55	0.55	0.52	0.49	0.49	0.50	0.53	0.49	0.49
	-0.75	0.56	0.52	0.52	0.54	0.55	0.46	0.49	0.54	0.45	0.51	0.53
	-1	0.48	0.38	0.48	0.42	0.46	0.35	0.40	0.45	0.40	0.45	0.46
	-1.25	0.39	0.36	0.43	0.39	0.37	0.32	0.31	0.39	0.35	0.36	0.38
	-1.5	0.36	0.34	0.36	0.33	0.28	0.30	0.26	0.30	0.29	0.28	0.30
	-1.75	0.33	0.29	0.31	0.29	0.26	0.24	0.23	0.27	0.28	0.26	0.26
	-2	0.32	0.28	0.28	0.26	0.26	0.22	0.22	0.24	0.26	0.25	0.25
	$\lambda \in L$	0.84	0.78	0.75	0.81	0.79	0.82	0.77	0.80	0.83	0.82	0.82
Normal	0	0.78	0.61	0.65	0.68	0.33	0.57	0.12	0.61	0.71	0.67	1.00
	-0.25	0.78	0.62	0.66	0.68	0.34	0.51	0.12	0.61	0.71	0.67	1.00
	-0.5	0.77	0.63	0.68	0.68	0.37	0.47	0.12	0.61	0.71	0.67	1.00
	-0.75	0.76	0.51	0.66	0.65	0.29	0.48	0.12	0.60	0.67	0.65	0.96
	-1	0.74	0.36	0.66	0.64	0.32	0.48	0.12	0.59	0.62	0.61	0.9
	-1.25	0.73	0.31	0.59	0.60	0.43	0.49	0.12	0.59	0.56	0.52	0.80
	-1.5	0.71	0.31	0.53	0.59	0.45	0.48	0.14	0.59	0.55	0.51	0.79
	-1.75	0.71	0.30	0.51	0.52	0.45	0.41	0.14	0.57	0.46	0.43	0.66
	-2	0.68	0.28	0.49	0.51	0.45	0.38	0.15	0.56	0.44	0.41	0.63
	$\lambda \in L$	0.78	0.67	0.93	0.69	0.77	0.69	0.18	0.61	0.71	0.67	1.00
Pareto	0	0.11	0.68	0.85	0.63	0.48	0.23	0.39	0.29	0.56	0.48	0.70
	-0.25	0.11	0.69	0.86	0.65	0.52	0.23	0.42	0.30	0.56	0.49	0.72
	-0.5	0.11	0.69	0.86	0.66	0.52	0.23	0.44	0.33	0.56	0.51	0.74
	-0.75	0.11	0.69	0.86	0.66	0.52	0.24	0.46	0.36	0.56	0.56	0.75
	-1	0.11	0.69	0.82	0.66	0.48	0.24	0.41	0.33	0.57	0.48	0.64
	-1.25	0.11	0.69	0.82	0.67	0.47	0.24	0.40	0.33	0.57	0.48	0.64
	-1.5	0.11	0.69	0.82	0.67	0.47	0.24	0.40	0.33	0.57	0.48	0.64
	-1.75	0.11	0.69	0.82	0.67	0.47	0.24	0.40	0.33	0.57	0.48	0.64
	-2	0.11	0.69	0.82	0.67	0.47	0.24	0.40	0.33	0.57	0.48	0.64
	$\lambda \in L$	0.11	0.69	0.88	0.67	0.53	0.25	0.46	0.39	0.60	0.58	0.78

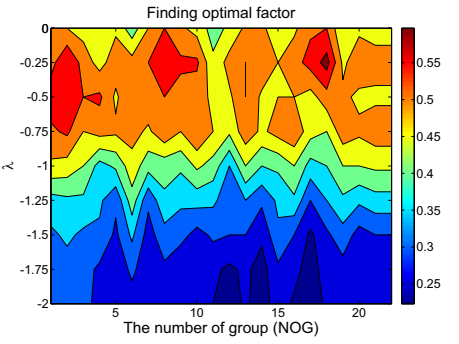
constraint will be 0. It means that there is often no solution. Thus the final results are acceptable.

5 Conclusions and Further Works

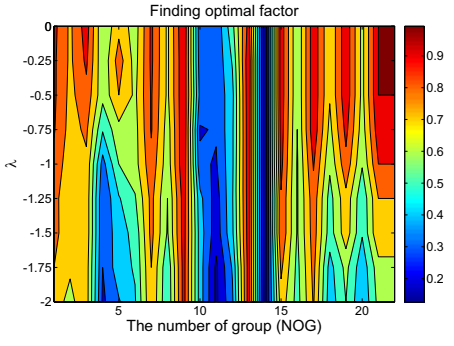
In some applications, there are a test cost constraint and common costs. The OSRT problem with common cost is challenging and interesting. In this paper, we proposed a heuristic algorithm to this problem. It produces satisfactory results on the Mushroom dataset. The parameter λ is employed in the penalty function such that expensive tests are more unlikely chosen. If λ influences the quality of the result significantly, the final result produced by the competition approach is generally better and stable. In the



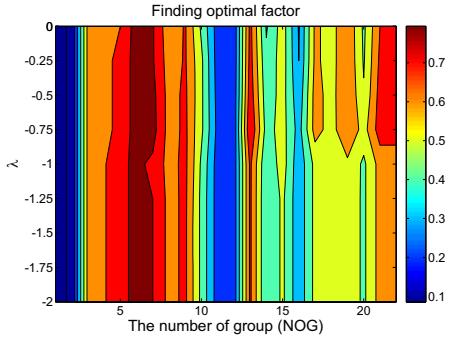
(a) Results by the competition approach



(b) Uniform distribution



(c) Normal distribution



(d) Bounded Pareto distribution

Fig. 1. The influence of NOG and different λ on FOF in three distributions in Mushroom dataset

future, we will study the influence of different common costs on the results. We will also develop more sophisticated algorithms to produce even better results.

Acknowledgments. This work is in part supported by National Science Foundation of China under Grant No. 60873077, 61170128, the Natural Science Foundation of Fujian Province, China under Grant No. 2011J01374, and the Education Department of Fujian Province under Grant No. JA11176. We would like to thank Dominik Ślęzak for valuable comments.

References

1. Chai, X., Deng, L., Yang, Q., Ling, C.X.: Test-cost sensitive naïve bayes classification. In: ICDM, pp. 51–58 (2004)
2. He, H., Min, F., Zhu, W.: Attribute reduction in test-cost-sensitive decision systems with common-test-costs. In: ICMLC, vol. 1, pp. 432–436 (2011)
3. Hu, Q., Yu, D., Liu, J., Wu, C.: Neighborhood rough set based heterogeneous feature subset selection. Information Sciences 178(18), 3577–3594 (2008)

4. Jensen, R., Shen, Q.: Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches. IEEE Transactions on Knowledge and Data Engineering 16(12), 1457–1471 (2004)
5. Kukar, M., Kononenko, I.: Cost-sensitive learning with neural networks. In: ECAI, pp. 445–449 (1998)
6. Min, F., He, H., Qian, Y., Zhu, W.: Test-cost-sensitive attribute reduction. Information Sciences 181, 4928–4942 (2011)
7. Min, F., Liu, Q.: A hierarchical model for test-cost-sensitive decision systems. Information Sciences 179, 2442–2452 (2009)
8. Min, F., Zhu, W.: Attribute reduction with test cost constraint. Journal of Electronic Science and Technology of China 9(2), 97–102 (2011)
9. Min, F., Zhu, W.: Minimal cost attribute reduction through backtracking. To appear in DTA (2011)
10. Min, F., Zhu, W.: Optimal sub-reducts in the dynamic environment. To appear in IEEE GrC (2011)
11. Min, F., Zhu, W.: Optimal Sub-Reducts with Test Cost Constraint. In: Yao, J., Ramanna, S., Wang, G., Suraj, Z. (eds.) RSKT 2011. LNCS(LNAI), vol. 6954, pp. 57–62. Springer, Heidelberg (2011)
12. Min, F., Zhu, W., Zhao, H., Pan, G.: Coser: Cost-senstive rough sets (2011), http://grc.fjzs.edu.cn/~fmin/coser/
13. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)
14. Qian, Y., Liang, J., Pedrycz, W., Dang, C.: Positive approximation: An accelerator for attribute reduction in rough set theory. Artificial Intelligence 174(9-10), 597–618 (2010)
15. Ślęzak, D.: Approximate entropy reducts. Fundamenta Informaticae 53(3-4), 365–390 (2002)
16. Turney, P.D.: Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. Journal of Artificial Intelligence Research 2, 369–409 (1995)
17. Wang, G.: Attribute Core of Decision Table. In: Alpigini, J.J., Peters, J.F., Skowron, A., Zhong, N. (eds.) RSCTC 2002. LNCS (LNAI), vol. 2475, pp. 213–217. Springer, Heidelberg (2002)
18. Yang, Q., Wu, X.: 10 challenging problems in data mining research. International Journal of Information Technology and Decision Making 5(4), 597–604 (2006)
19. Yao, Y.: A Partition Model of Granular Computing. In: Peters, J.F., Skowron, A., Grzymała-Busse, J.W., Kostek, B.z., Świniarski, R.W., Szczuka, M.S. (eds.) Transactions on Rough Sets I. LNCS, vol. 3100, pp. 232–253. Springer, Heidelberg (2004)
20. Yao, Y., Wong, S.: A decision theoretic framework for approximating concepts. International Journal of Man-machine Studies 37, 793–809 (1992)
21. Yao, Y., Zhao, Y.: Attribute reduction in decision-theoretic rough set models. Information Sciences 178(17), 3356–3373 (2008)
22. Zhang, W., Mi, J., Wu, W.: Knowledge reductions in inconsistent information systems. Chinese Journal of Computers 26(1), 12–18 (2003)
23. Zhou, Z., Liu, X.: Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering 18(1), 63–77 (2006)
24. Zhu, W.: Topological approaches to covering rough sets. Information Sciences 177(6), 1499–1508 (2007)
25. Zhu, W., Wang, F.: Reduction and axiomization of covering generalized rough sets. Information Sciences 152(1), 217–230 (2003)
26. Ziarko, W.: Variable precision rough set model. Journal of Computer and System Sciences 46(1), 39–59 (1993)

Ensembles of Bireducts: Towards Robust Classification and Simple Representation*

Dominik Ślęzak[1,2] and Andrzej Janusz[1]

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

Abstract. We introduce the notion of a bireduct, which is an extension of the notion of a reduct developed within the theory of rough sets. For a decision system $\mathbb{A} = (U, A \cup \{d\})$, a bireduct is a pair (B, X), where $B \subseteq A$ is a subset of attributes that discerns all pairs of objects in $X \subseteq U$ with different values of the decision attribute d, and where B and X cannot be, respectively, reduced and extended without losing this property. We investigate the ability of ensembles of bireducts (B, X) characterized by significant diversity with respect to both B and X to represent knowledge hidden in data and to serve as the means for learning robust classification systems. We show fundamental properties of bireducts and provide algorithms aimed at searching for ensembles of bireducts in data. We also report results obtained for some benchmark data sets.

Keywords: Attribute Subset Selection, Inexact Dependencies, Classifier Ensembles, Discernibility, Decision Rules, Randomized Search.

1 Introduction

Attribute subset selection plays an important role in knowledge discovery [9]. It establishes the basis for more efficient classification, prediction and approximation models. It also provides the users with a better insight into data dependencies. In this paper, we concentrate on attribute subset selection methods originating from the theory of rough sets [12]. There are numerous rough-set-based algorithms aimed at searching for so called *reducts* – irreducible subsets of attributes that satisfy predefined criteria for keeping enough information about decisions. Those criteria are verified on the training data and, usually, they encode more or less directly the risk of misclassification by if-then decision rules with their antecedents referring to the values of investigated attribute subsets and their consequents referring to decisions.

* The authors were supported by the grant N N516 077837 from the Ministry of Science and Higher Education of the Republic of Poland and by the National Centre for Research and Development (NCBiR) under the grant SP/I/1/77065/10 by the Strategic scientific research and experimental development program: "Interdisciplinary System for Interactive Scientific and Scientific-Technical Information".

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 64–77, 2011.

Original definition of a reduct is quite restrictive, requiring that it should determine decisions or, if data inconsistencies do not allow full determinism, provide the same level of information about decisions as the complete set of attributes. There are a number of approaches to formulate and search for approximate or inexact reducts, which *almost* preserve the decision information (see [14,15] for some examples). Approximate reducts are usually smaller than standard reducts, providing the basis for learning more efficient classifiers. There are, however, the following issues to be addressed.

First of all, there is a variety of methods of searching for approximate reducts in decision systems. The criteria usually include formulas for functions measuring degrees of decision information induced by subsets of attributes and thresholds for those functions' values specifying which subsets of attributes are *good enough*. The choice of functions may depend on the nature of particular data sets and methods of learning classifiers based on reduced sets of attributes. The approximation thresholds need to be carefully tuned. This fact often makes the whole process of attribute subset selection quite unclear to the users.

The second issue is related to a popular idea of building classifier ensembles [2]. Combining classifiers is efficient especially if they are different from each other, where differences can be understood in many ways. For instance, one may focus on ensembles of classifiers learned from reducts that include as different attributes as possible. In this way, one may increase stability of the classification and improve the ability to represent data dependencies to the users. Unfortunately, the current approximate reduct criteria do not allow controlling which parts of data are problematic for particular reducts. For example, when building an ensemble where individual reducts are supposed to correctly classify at least 90% of the training objects, we may not anticipate that each of the resulting classifiers will have problems with the same 10% of objects.

Given the above challenges, we propose a new extension of the original rough-set-based notion of a reduct, whose interpretation is simpler than for various types of approximate reducts known from the literature. In our approach, the emphasis is on both, a subset of attributes that describes the decision classes and a subset of objects for which such a description is valid. Inspired by the methodology of biclustering [10], where it is crucial to work with objects and attributes simultaneously, we refer to the proposed notion as a *bireduct*.

The remainder of the paper is organized as follows: In Section 2, we define information bireducts and decision bireducts for an information system $\mathbb{A} = (U, A)$ and a decision system $\mathbb{A} = (U, A \cup \{d\})$, respectively. They take a form of pairs (B, X), where subsets $B \subseteq A$ and $X \subseteq U$ are linked together by means of standard rough-set-based discernibility [12]. In Section 3, we discuss the algorithmic framework for searching for bireducts, following some earlier works on the order-based genetic algorithms searching for standard reducts [1] and our own research on permutation-based search for approximate reducts [14]. In Section 4, we investigate properties of the bireduct-based classifier ensembles for some benchmark data sets. We also measure the overlaps of subsets of attributes and objects for bireducts in the ensembles. In Section 5, we conclude the paper.

	Outlook	Temp.	Humid.	Wind	Sport?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cold	Normal	Weak	Yes
6	Rain	Cold	Normal	Strong	No
7	Overcast	Cold	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cold	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

({O,T,W},{1-14})
({O,W},{1-8,10,12-14})
({O,T},{1-3,5,7-9,12-14})
({O,H},{1-5,7-13})
({O,W},{3-7,9-14})
({W},{2-6,9-10,13-14})
({O,H,W},{1-14})
({T,W},{2-3,5-6,8-9,13-14})
({H},{3-5,7,9-13})
({T,W},{1-2,4-5,7,9-10,14})
({T,H,W},{2-3,5,7-13})
({H,T},{1-2,4-5,7,9-13})
({O},{1-5,7-8,10,12-13})
({H,W},{1-2,5-6,8-10,13-14})
({T},{1-2,4,6,10-12})

Fig. 1. Left: Decision system $\mathbb{A} = (U, A \cup \{d\})$ with 14 objects in U, four attributes in A, and $d =$ Sport?. **Right**: Examples of decision bireducts for $\mathbb{A}$.

2 Theoretical Foundations

We use the standard notation of information systems to represent data [12]. By an information system we mean a tuple $\mathbb{A} = (U, A)$, where U is a set of objects and A is a set of attributes. For simplicity, we refer to the elements of U using their ordinal numbers $i = 1, ..., |U|$, where $|U|$ denotes the cardinality of U. We treat attributes $a \in A$ as functions $a : U \to V_a$, V_a denoting a's domain. By a decision system we mean $\mathbb{A} = (U, A \cup \{d\})$, where $d \notin A$ is a distinguished decision attribute. The values $v_d \in V_d$ correspond to decision classes that we want to describe using the values of attributes in A. Thus, decision systems can be employed within the standard supervised learning framework.

Let us consider the notion of a reduct, which is one of the most important contributions of the theory of rough sets into knowledge discovery and data mining. We say that $B \subseteq A$ is an information reduct for $\mathbb{A} = (U, A)$, if and only if it is an irreducible subset of attributes such that each pair of objects $i, j \in U$, which is discerned by A (i.e. there exists $a \in A$ such that $a(i) \neq a(j)$), is also discerned by B. Going further, we say that $B \subseteq A$ is a decision reduct for $\mathbb{A} = (U, A \cup \{d\})$, if and only if it is an irreducible subset of attributes such that each pair $i, j \in U$ satisfying inequality $d(i) \neq d(j)$ is discerned by B. For the decision system in Figure 1, there are two decision reducts: $\{\text{Outlook}, \text{Temperature}, \text{Wind}\}$ and $\{\text{Outlook}, \text{Humidity}, \text{Wind}\}$ (or $\{\text{O}, \text{T}, \text{W}\}$ and $\{\text{O}, \text{H}, \text{W}\}$ for short).

Definition 1. *Let $\mathbb{A} = (U, A)$ be an information system. A pair (B, X), where $B \subseteq A$ and $X \subseteq U$, is called an information bireduct, if and only if B discerns all pairs of objects in X (i.e. for each $i, j \in X$ there exists $b \in B$ such that $b(i) \neq b(j)$), and the following properties hold:*

1. *There is no proper subset $C \subsetneq B$ such that C discerns all pairs in X;*
2. *There is no proper superset $Y \supsetneq X$ such that B discerns all pairs in Y.*

It is interesting to compare information bireducts with templates studied in the association rule mining (cf. [11]) or concepts known from the formal concept analysis [5]. Templates aim at describing maximum amounts of objects with the same (or similar enough) values on maximum amounts of attributes. Similarly, concepts are defined as non-extendable subsets of objects that are indiscernible with respect to non-extendable subsets of attributes. On the other hand, information bireducts describe non-extendable subsets of objects that are discernible using irreducible subsets of attributes. We may say that templates and concepts correspond to the most regular areas of data, while information bireducts – to the most irregular (one might even claim – the most informative) areas. Hence, information bireducts could be also called anti-templates or anti-concepts. For a preliminary example of their application, we refer the reader to [7].

Definition 2. *Let $\mathbb{A} = (U, A \cup \{d\})$ be a decision system. A pair (B, X), where $B \subseteq A$ and $X \subseteq U$, is called a decision bireduct, if and only if B discerns all pairs $i, j \in X$ where $d(i) \neq d(j)$, and the following properties hold:*

1. *There is no $C \subsetneq B$ such that C discerns all pairs $i, j \in X$ where $d(i) \neq d(j)$;*
2. *There is no $Y \supsetneq X$ such that B discerns all pairs $i, j \in Y$ where $d(i) \neq d(j)$.*

A decision bireduct (B, X) can be regarded as an inexact functional dependence linking the subset of attributes B with the decision d in a degree X, denoted by $B \Rrightarrow_X d$. The objects in $U \setminus X$ can be treated as the outliers. The objects in X can be used to learn a classifier based on B from data. For instance, one can partition X with respect to its elements' values on B and treat the combinations of values labeling partition classes (called indiscernibility classes in the rough set literature [12]) as the antecedents of rules pointing at specific decision values, uniquely defined within X thanks to the properties of decision bireducts.

Figure 1 shows a few decision bireducts for a well-known example of a decision system. The number of all decision bireducts for this data set is far higher than illustrated. One may notice that the same $B \subseteq A$ can occur as a component of many bireducts, with different subsets of objects. Just like in case of standard reducts [11], one can provide a Boolean interpretation of bireducts as prime implicants of an appropriately constructed CNF formula encoding the data.

Proposition 1. *Let $\mathbb{A} = (U, A \cup \{d\})$ be a decision system. Consider the following Boolean formula with variables $\overline{i}$, $i = 1, ..., |U|$, and $\overline{a}$, $a \in A$:*

$$\tau_{\mathbb{A}}^{bi} = \bigwedge_{i,j:\, d(i) \neq d(j)} \left(\overline{i} \vee \overline{j} \vee \bigvee_{a:\, a(i) \neq a(j)} \overline{a} \right). \tag{1}$$

An arbitrary pair (B, X), $B \subseteq A$, $X \subseteq U$, is a decision bireduct, if and only if the Boolean formula $\bigwedge_{a \in B} \overline{a} \wedge \bigwedge_{i \notin X} \overline{i}$ is the prime implicant for $\tau_{\mathbb{A}}^{bi}$.

The analogous result can be formulated for information bireducts. In both cases, the proofs are straightforward. According to Proposition 1, it may be convenient to describe decision bireducts in terms of their attributes and outliers, with an implicit assumption that the most meaningful bireducts shall minimize both those factors. Proposition 1 also emphasizes that the number of bireducts is usually far higher than the number of standard reducts. For instance, as already mentioned, there are only two standard decision reducts for the decision system illustrated by Figure 1. On the other hand, the family of all decision bireducts can be represented by the following Boolean derivation, where the CNF form encodes the pairs of objects with different decision values (such as 1 and 3 that differ on attribute O, or 1 and 4 that differ on O and T) and the DNF form encodes decision bireducts (such as bireduct with attributes O, T, W and no outliers, or bireduct with attributes O, T and outliers 4, 6, 10, 11).

$$\begin{aligned} \tau^{bi}_{\mathbb{A}} &\equiv (\overline{1} \vee \overline{3} \vee \overline{\mathrm{O}}) \wedge (\overline{1} \vee \overline{4} \vee \overline{\mathrm{O}} \vee \overline{\mathrm{T}}) \wedge ... \wedge (\overline{13} \vee \overline{14} \vee \overline{\mathrm{O}} \vee \overline{\mathrm{T}} \vee \overline{\mathrm{H}} \vee \overline{\mathrm{W}}) \\ &\equiv (\overline{\mathrm{O}} \wedge \overline{\mathrm{T}} \wedge \overline{\mathrm{W}}) \vee (\overline{\mathrm{O}} \wedge \overline{\mathrm{W}} \wedge \overline{9} \wedge \overline{11}) \vee (\overline{\mathrm{O}} \wedge \overline{\mathrm{T}} \wedge \overline{4} \wedge \overline{6} \wedge \overline{10} \wedge \overline{11}) \vee ... \end{aligned} \tag{2}$$

3 Algorithmic Foundations

There are numerous results on NP-hardness of the search for optimal attribute subsets in data. In a case of rough-set-based reducts and their approximate extensions, optimization usually aims at decreasing the number of used attributes or the complexity of the corresponding rule-set-based classifiers (see [14]). With the bireducts, as discussed in Section 2, this approach may be generalized towards simultaneous minimization of the involved attributes and the resulting outliers. Intuitively, the related optimization problems are expected to be NP-hard as well, although the proofs of the corresponding mathematical results depend on a way of balancing between the amounts of attributes and outliers. In this paper, we skip this part of analysis. Instead, we focus on an algorithmic mechanism that searches for bireducts with their corresponding optimization principles taken into account implicitly rather than explicitly.

Algorithm 1 takes as an input a permutation $\sigma : \{1, ..., n+m\} \rightarrow \{1, ..., n+m\}$ mixing the ordinal numbers of attributes counted from 1 to n, $n = |A|$, together with objects represented by numbers from $n+1$ to $n+m$, $m = |U|$. The algorithm is initiated with the pair (B, X), where $B = A$ and $X = \emptyset$. Then, it examines the values of $\sigma(i)$, for $i = 1, ..., n+m$. Depending on whether $\sigma(i)$ corresponds to an attribute (the case of $\sigma(i) \leq n$) or an object (the case of $\sigma(i) > n$; the corresponding object is then retrieved as $\sigma(i) - n$), it attempts to remove it from B or to add it to X, respectively. The removal/addition conditions are defined as $B \setminus \{a_{\sigma(i)}\} \Rrightarrow_X d$ and $B \Rrightarrow_{X \cup \{\sigma(i)-n\}} d$, using the inexact functional dependence notation introduced in Section 2.

Proposition 2. *Let decision system $\mathbb{A} = (U, A \cup \{d\})$ be given. For each permutation $\sigma : \{1, ..., n+m\} \rightarrow \{1, ..., n+m\}$, where $n = |A|$ and $m = |U|$, the output (B, X) of Algorithm 1 is a decision bireduct.*

Algorithm 1. Decision bireduct calculation for $\mathbb{A} = (U, A \cup \{d\})$

Input: $\sigma : \{1, ..., n+m\} \to \{1, ..., n+m\}$, $n = |A|$, $m = |U|$
Output: (B, X), $B \subseteq A$, $X \subseteq U$

```
B ← A, X ← ∅
for i = 1 to n + m do
  if σ(i) ≤ n then
    if B \ {a_σ(i)} ⇛_X d then
      B ← B \ {a_σ(i)}
    end if
  else
    if B ⇛_{X ∪ {σ(i)−n}} d then
      X ← X ∪ {σ(i) − n}
    end if
  end if
end for
return (B, X)
```

A similar result, for an analogously designed algorithm, can be formulated for the information bireducts. The proofs are straightforward. In fact, Algorithm 1 is a generalization of a permutation-based method developed for searching for standard rough-set-based reducts (see [1]), which was also adapted for the approximate reducts (see [14]). The difference is that in case of the standard reducts the algorithm works with permutations $\sigma : \{1, ..., n\} \to \{1, ..., n\}$, $n = |A|$, and needs to assure irreducibility of the generated subsets of attributes. In the case of Algorithm 1 we need to assure both, irreducibility of the subset of attributes and non-extendability of the subset of objects.

It is already known that in the case of the original permutation-based algorithm, the standard (or approximate) reducts that contain less common attributes are more likely to be obtained by employing a randomly chosen $\sigma : \{1, ..., n\} \to \{1, ..., n\}$ [14]. This property has an impact on the search for the most meaningful reducts, as well as the reduct ensembles – the sets of reducts with minimal intersections [15]. The analogous behavior can be seen for Algorithm 1, where randomly chosen permutations $\sigma : \{1, ..., n+m\} \to \{1, ..., n+m\}$ are more likely to lead to the bireducts with more diverse attributes and outliers. As it was mentioned in the discussion in Section 1, such an ability to control the ensembles of bireducts with respect to the areas of objects that they cover is especially important for robustness of the resulting classification systems and completeness of data representation.

An additional question arises for the subsets of attributes occurring in multiple bireducts (see Section 2). Assume that a randomly chosen permutation leads to a bireduct with a specific $B \subseteq A$. The question is: Which $X \subseteq U$ is the most probable to be obtained together with B? An intuitive answer would be to consider the largest X such that $B \Rrightarrow_X d$. The size of such X corresponds to the criterion used for so called M-approximate reducts (see [15]). However, it is clear that the bireducts lead towards a significantly different framework.

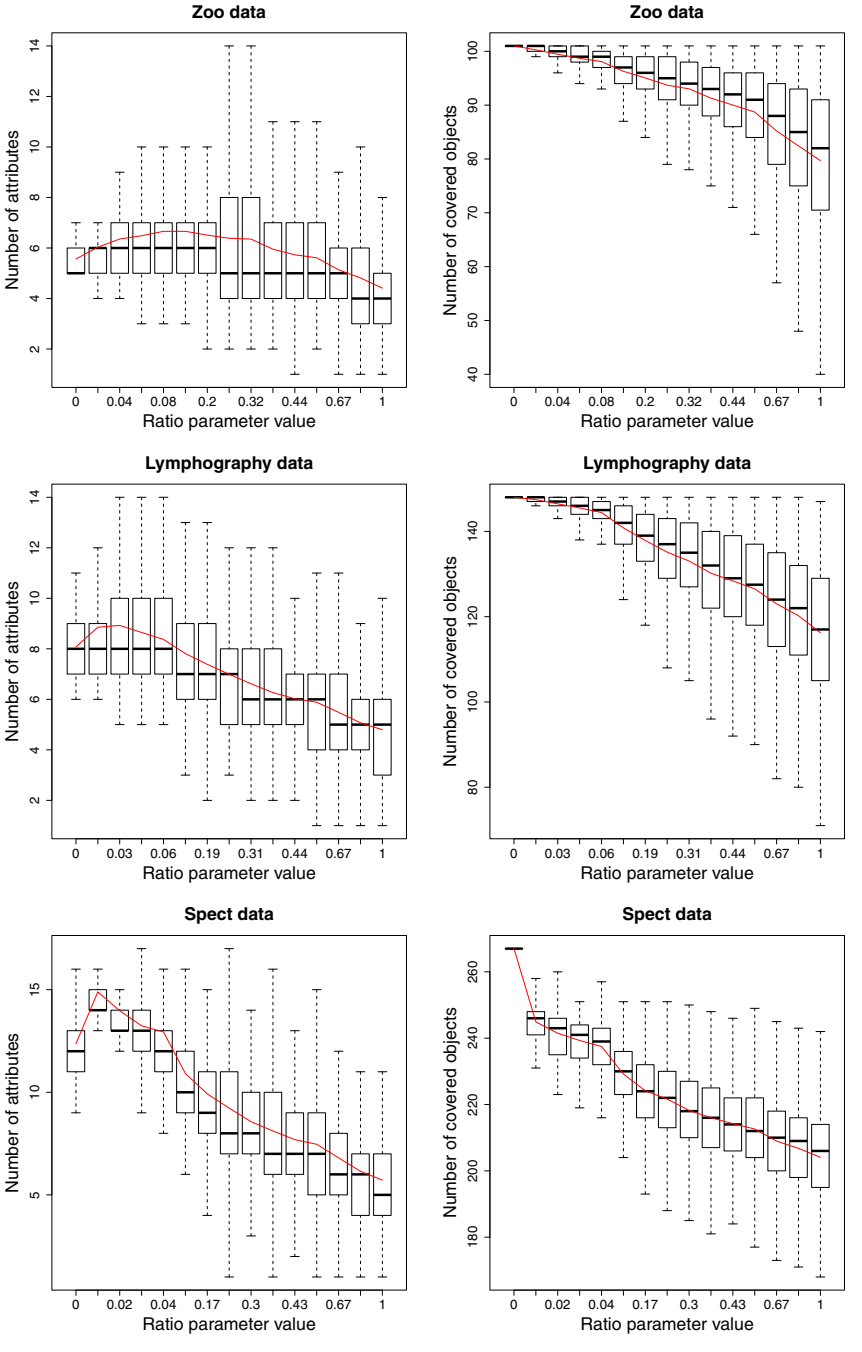


Fig. 2. Statistics (quantiles and the average) for number of attributes (left) and objects (right) for bireducts of three benchmark data sets

4 Experimental Evaluation

We conducted a series of experiments in order to investigate usefulness of the decision bireduct ensembles for classification purposes. All the experiments were run within the R System [13]. We utilized three popular benchmark data sets from the UCI repository [3]: *zoo*, *lymphography* and *spect*.

Even for relatively small data sets, it is impossible to examine all permutations. This is why the analysis of the outcomes of randomly chosen permutations is so important (see Section 3). There is a variety of techniques of generating permutations $\sigma : \{1, ..., n\} \rightarrow \{1, ..., n\}$ yielding standard or approximate reducts, spanning from advanced genetic algorithms (see [1] for references) to relatively simple generators whose performance depends on the convergence of randomly chosen permutations towards potentially optimal outputs. In this paper, we follow the latter approach. We search for decision bireduct ensembles by feeding Algorithm 1 with fully randomly chosen permutations $\sigma : \{1, ..., n + m\} \rightarrow \{1, ..., n + m\}$ and memorizing a fixed number of output bireducts (B, X).

The process of generating a bireduct ensemble with desired properties (such as the average numbers of attributes and outliers or more sophisticated measures introduced later in this section) can be influenced by the way that the permutations are generated. A special case is when all objects are at the beginning of $\sigma : \{1, ..., n + m\} \rightarrow \{1, ..., n + m\}$ – then, after m steps of Algorithm 1, X becomes equal to U (or almost U, in case there are some inconsistencies in $\mathbb{A}$) and the process becomes similar to the search for standard reducts. A modification of the process leads to bireducts (B, X) with X being a bit smaller than U and B being significantly smaller than A – it is enough to increase probability that some attributes will occur closer to beginning of a permutation sequence. Let us introduce a *ratio* parameter that corresponds to the weight of attributes during the permutation construction – the higher the weight, the more attributes appear early in the sequence. The neutral ratio, i.e., the one yielding uniform distribution of attributes and objects within permutations, equals to $|U|/|A|$. In this paper, we investigate the ratios spanning from 0 to $2|U|/|A|$. For convenient comparison of results between different data sets, we additionally scale the ratios from $[0, 2|U|/|A|]$ to the $[0, 1]$ interval.

The first step of our experimental analysis is to compare decision bireducts with standard decision reducts in terms of their size. For each data set we computed 1000 decision reducts and 14000 bireducts for 14 different values of the ratio parameter (1000 bireducts for each ratio value). Figure 2 summarizes the number of involved attributes (left side) and objects (right side) for different ratio values. Ratio equal to 0 indicates the standard decision reducts.

As expected, the average number of objects covered by a bireduct (the solid line in the plots on the right) drops when the ratio is increased. One can also notice that the spreads of the number of objects increase with the increasing ratios. However, these observations do not hold for the average number of attributes (the solid line in the plots on the left), which slightly increases for small ratio values and then drops below the average for the standard reducts.

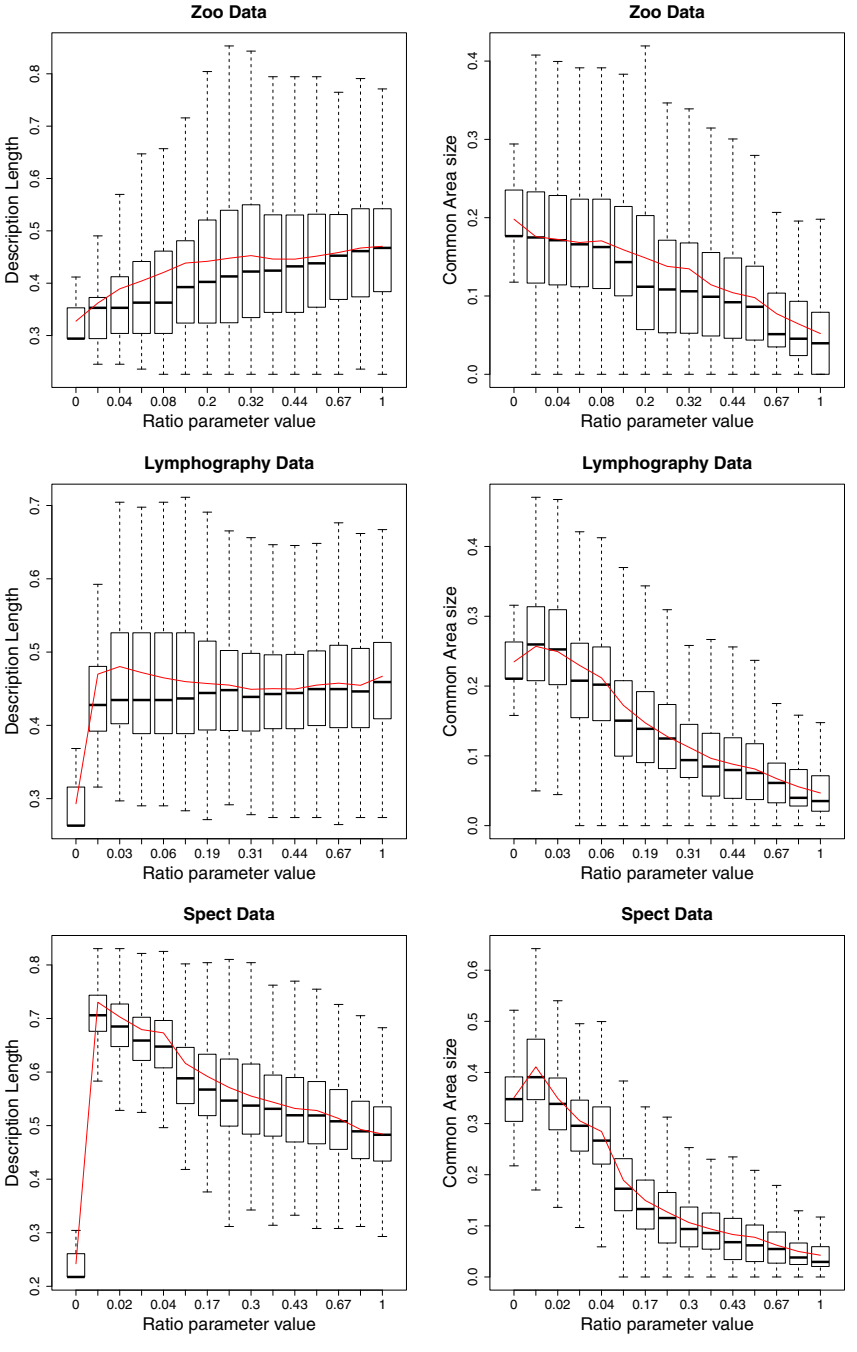


Fig. 3. Statistics (quantiles and the average) for description length (left) and intersection size (right) for bireducts of three benchmark data sets

To further investigate the relation between the number of attributes and objects in bireducts we examined two additional properties. The first one is the reduct description length defined as

$$DescLength((B, X)) = |B| / |A| + |U \setminus X| / |U| . \quad (3)$$

The description length of a bireduct can be interpreted as the length of the Boolean formula representing the corresponding prime implicant by means of the number of attributes $|B|$ and the number of outliers $|U \setminus X|$, additionally normalized by $|A|$ and $|U|$, respectively.

The second property relates to the overlap of two bireducts, defined as

$$OverlapSize((B_1, X_1), (B_2, X_2)) = |B_1 \cap B_2| / |A| \times |X_1 \cap X_2| / |U| . \quad (4)$$

The average overlap size is an indicator of a bireduct ensemble diversity. Intuitively, if for a given set of bireducts the average overlap size is small, this set is more likely to cover broader part of U. Figure 3 presents statistics for these two properties, computed for different ratio values.

We have also investigated the influence of the ratio parameter and classifier aggregation methods on the classification results. We performed ten 5-fold cross-validation tests for each data set and each ratio value. In a single training/testing cycle, 1000 bireducts were constructed for every ratio value using the training data. Rule sets corresponding to indiscernibility classes within bireducts were used as classifiers. Decision classes of objects from the test set were predicted using two aggregation methods – the majority voting and the balanced support weighted voting. Such weighting strategies are quite popular in the machine learning literature (see [15]), although we should emphasize that in case of bireducts (B, X) they are computed with respect to X instead of U.

Because decision classes in the utilized data sets were significantly imbalanced, we used two quality measures to evaluate our classifiers – the mean accuracy and the balanced accuracy. The mean accuracy is simply a percentage of correctly classified objects. The balanced accuracy is the mean of a percentage of correctly classified objects within each decision class (see e.g. [16]). This measure is insensitive to imbalanced class distribution. The balanced accuracy gives more weight to instances from minority classes, whereas the mean accuracy treats all objects alike and usually favors the majority class.

The two different aggregation methods that we have used aim at maximizing different quality measures. The majority voting scheme classifies a test object to the decision class indicated by the highest number of triggered rules derived from bireducts in the ensemble. This voting scheme can be biased toward larger decision classes and usually favors the mean accuracy measure.

The balanced support weighted voting scheme weights each vote using the support of the corresponding rule. The class of an object then is decided by taking into consideration distribution of decisions in the training data. This method is preferable for maximizing the balanced accuracy. Figures 4 and 5 show the scores in terms of the mean and the balanced accuracies of the majority (on the left) and the balanced support weighted (on the right) voting schemes.

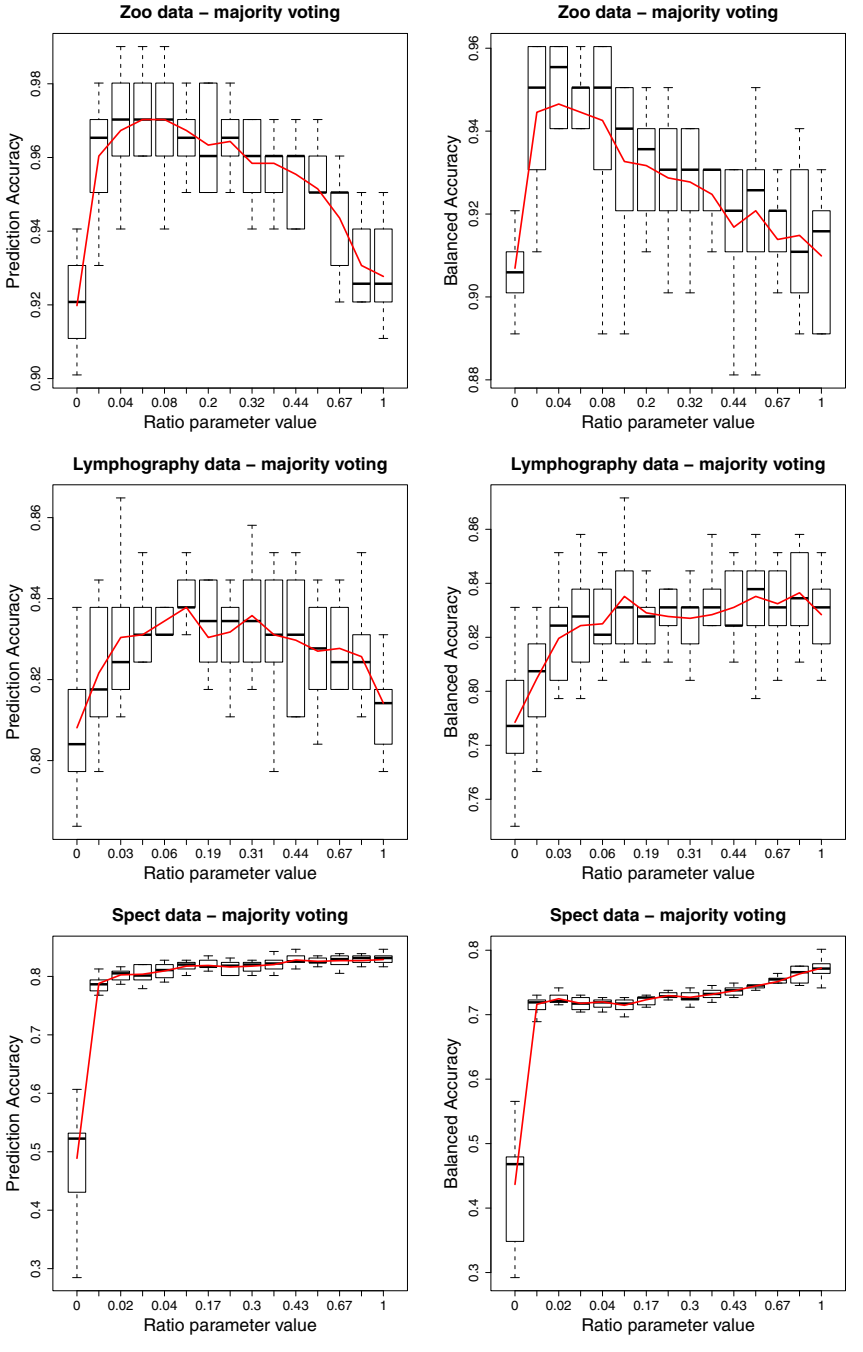


Fig. 4. Classification accuracies (left) and balanced accuracies (right) of bireduct ensembles acquired by the majority voting

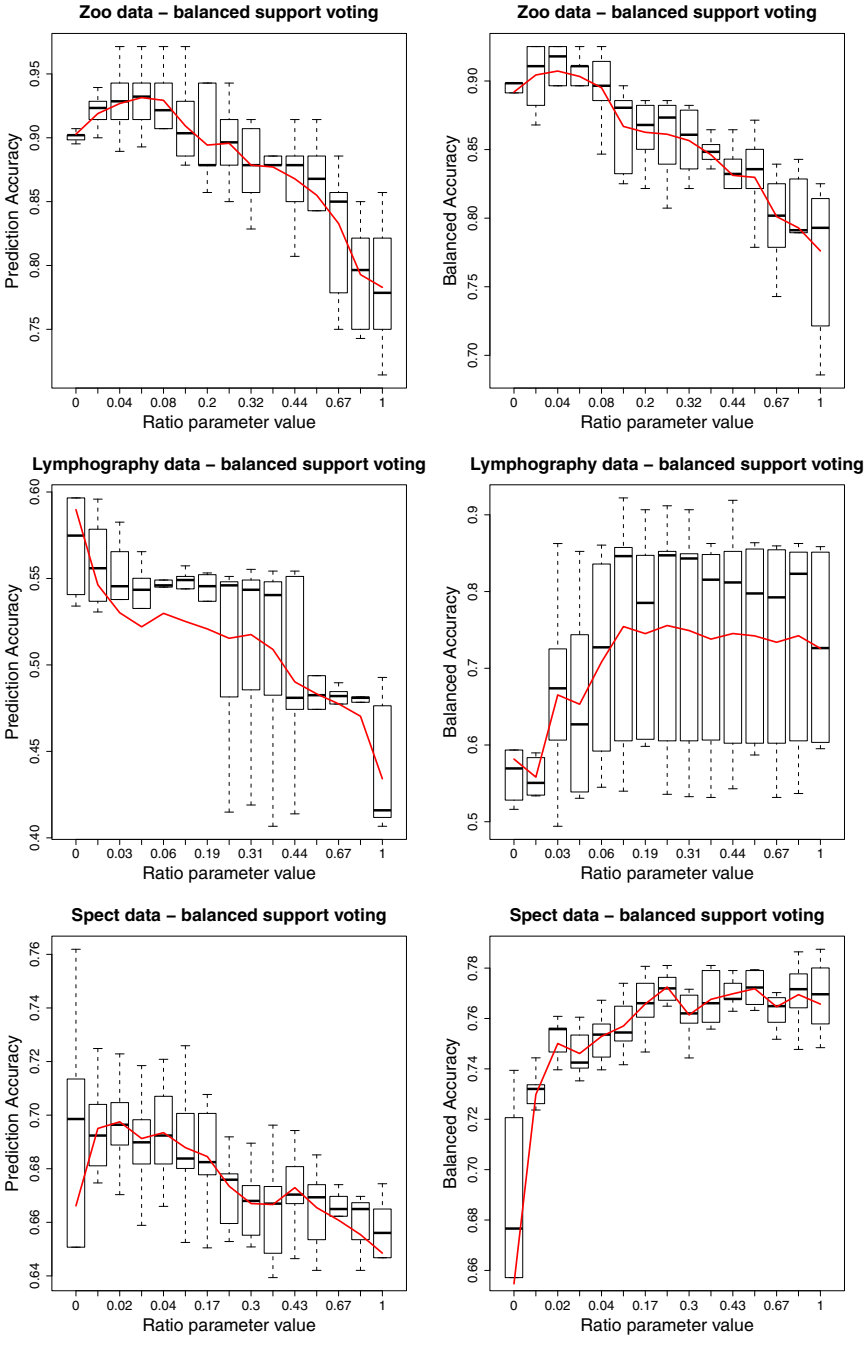


Fig. 5. Classification accuracies (left) and balanced accuracies (right) of bireduct ensembles acquired by the balanced support weighted voting

The results clearly show strengths of the bireduct ensembles. Their scores on all three data sets are comparable with the best results in the machine learning literature [4,8]. For small values of the ratio parameter, the bireduct classifiers consistently outperformed the ensembles of standard reducts (the ratio value set to 0) regardless of the aggregation method and the quality measure used. The most significant difference is visible on the *spect* data where for some ratio values the average scores of bireduct ensembles are greater by 0.30.

The improvement over the standard reducts is also noticeable with respect to the number of unrecognized objects. This factor for the standard reduct ensembles was several orders of magnitude greater than for any of the considered bireduct ensembles. For example, on the *spect* data, the standard reducts were not able to recognize over 35% of objects, whereas the bireduct ensemble computed for the ratio ≈ 0.01, on average, did not recognize only $\approx 0.006\%$ of objects. The number of objects unrecognized by the ensemble of bireducts drops to zero for ratios greater than 0.02, regardless of a data set.

5 Conclusions

We proposed a novel approach to construction of classifier ensembles based on the notion of a bireduct. This simple, yet having solid theoretical foundations concept allows to capture complex dependencies in data, which are difficult to express using other methods. Although we only examined the utilization of bireduct ensembles in the supervised classification framework, this notion can be easily put to use in other application types (see e.g. [7]).

The experiments show that application of bireducts helps in finding diverse sets of classification rules, which in turn increases the ensembles' performance. A neutral value of the ratio parameter (that does not interfere with natural proportions between objects and attributes in data; see Section 4) produces reasonable results for all the tested data sets. It suggests stability of the method. We also believe that tuning the ratio value may be more efficient than tuning parameters in other attribute subset selection approaches.

In future, we will further examine the properties of bireducts and compare them with those of standard rough-set-based reducts, as well as with, e.g., M-approximate reducts mentioned in Section 3. We will also investigate bireducts based on other types of discernibility including, e.g., some similarity-based extensions [6]. Last but not least, it will be important to verify how decision bireducts should be integrated with various classification models, especially those frequently used in the ensemble-based classification systems.

References

1. Bazan, J., Nguyen, H., Nguyen, S., Synak, P., Wróblewski, J.: Rough Set Algorithms in Classification Problem. In: Polkowski, L., Tsumoto, S., Lin, T. (eds.) Rough Set Methods and Applications. STUDFUZZ, vol. 56, pp. 49–88. Physica Verlag (2000)

2. Dietterich, T.G.: An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Machine Learning 40(2), 139–157 (2000)
3. Frank, A., Asuncion, A.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2010)
4. Frank, E., Kramer, S.: Ensembles of Nested Dichotomies for Multi-class Problems. In: Proc. of Int. Conf. on Machine Learning (ICML). ACM International Conference Proceeding Series, vol. 69 (2004)
5. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer (1998)
6. Janusz, A.: Similarity Relation in Classification Problems. In: Chan, C.-C., Grzymala-Busse, J.W., Ziarko, W.P. (eds.) RSCTC 2008. LNCS (LNAI), vol. 5306, pp. 211–222. Springer, Heidelberg (2008)
7. Janusz, A., Ślęzak, D.: An Unsupervised Model for Rule-based Similarity Learning from Textual Data: A General Idea. In: Proc. of Int. Workshop on Concurrency, Specification, and Programming (CS&P), pp. 229–237 (2011)
8. Kurgan, Ł.A., Cios, K.J., Tadeusiewicz, R., Ogiela, M.R., Goodenday, L.S.: Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis. Artificial Intelligence in Medicine 23(2), 149–169 (2001)
9. Liu, H., Motoda, H. (eds.): Computational Methods of Feature Selection. Chapman & Hall/CRC (2008)
10. Mirkin, B.: Mathematical Classification and Clustering. Kluwer (1996)
11. Nguyen, H.S.: Approximate Boolean Reasoning: Foundations and Applications in Data Mining. In: Peters, J.F., Skowron, A. (eds.) Transactions on Rough Sets V. LNCS, vol. 4100, pp. 334–506. Springer, Heidelberg (2006)
12. Pawlak, Z., Skowron, A.: Rudiments of Rough Sets. Information Sciences 177(1), 3–27 (2007)
13. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2008), http://www.R-project.org
14. Ślęzak, D.: Rough Sets and Functional Dependencies in Data: Foundations of Association Reducts. In: Gavrilova, M.L., Tan, C.J.K., Wang, Y., Chan, K.C.C. (eds.) Transactions on Computational Science V. LNCS, vol. 5540, pp. 182–205. Springer, Heidelberg (2009)
15. Ślęzak, D., Widz, S.: Is It Important Which Rough-Set-Based Classifier Extraction and Voting Criteria Are Applied Together? In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) RSCTC 2010. LNCS (LNAI), vol. 6086, pp. 187–196. Springer, Heidelberg (2010)
16. Wojnarski, M., Janusz, A., Nguyen, H.S., Bazan, J., Luo, C., Chen, Z., Hu, F., Wang, G., Guan, L., Luo, H., Gao, J., Shen, Y., Nikulin, V., Huang, T.-H., McLachlan, G.J., Bošnjak, M., Gamberger, D.: RSCTC'2010 Discovery Challenge: Mining DNA Microarray Data for Medical Diagnosis and Treatment. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) RSCTC 2010. LNCS (LNAI), vol. 6086, pp. 4–19. Springer, Heidelberg (2010)

Efficient Implementation of Recursive Queries in Major Object Relational Mapping Systems*

Aneta Szumowska, Marta Burzańska, Piotr Wiśniewski, and Krzysztof Stencel

Faculty of Mathematics and Computer Science
Nicolaus Copernicus University
Toruń, Poland
{iriz,quintria,pikonrad,stencel}@mat.uni.torun.pl

Abstract. The following paper presents the effects of combining two technologies: object-relational mapping and SQL's recursive queries. Both technologies are widely used in modern software, and yet, modern ORM systems still lack the support for recursive database querying. The currently used methods for querying graph and hierarchical structures are either inefficient, difficult to maintain, or do not allow for any portability. The authors of the following paper propose extensions to the general functionality of modern ORM systems and present the results for two ORM systems: Hibernate for Java and Django-models for Python. With this extension programmers using one of those systems can benefit from the support for the recursive queries offered by various object-relational database management systems and write a maintainable code compliant with the used ORM standard. The proposed solution works with IBM DB2, Oracle and PostgreSQL DBMS and proved to be many times faster than the approaches currently used.

1 Introduction

Recursive queries are a part of the SQL standard since 1999. During the last 10 years they gained much popularity and are now implemented in most of the relational database management systems. The basic form of a recursive query is an extension to the common table expressions (recursive CTE, RCTE).

Each RCTE consists of three parts: seed query, recursive query comprising references to the CTE being defined and an outer query that utilizes recursively generated data. Each of those queries may have their own selection predicates and have different tables declared in the FROM clause. Such expression is presented on listing 1.1 further in this paper.

More about history of recursive queries, their availability and efficiency of evaluation in modern DBMSs may be found in [1,2]. Despite the fact that recursive queries appeared quite a long time ago, intensive research is still being

* The authors were supported by the grant N N516 077837 from the Ministry of Science and Higher Education of the Republic of Poland and by the National Centre for Research and Development (NCBiR) under the grant SP/I/1/77065/10 by the Strategic scientific research and experimental development program: "Interdisciplinary System for Interactive Scientific and Scientific-Technical Information".

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 78–89, 2011.

conducted on their optimization [3,4]. The time that has passed allowed the field of recursive queries to mature and now they become increasingly popular among software developers.

When looking at the recent history of software development, one may observe that while the relational databases were (and still are) the main choice for data storage, the software developers moved toward object-oriented programming languages. From the attempts to combine both worlds in one software project emerged a number of problems collectively called the impedance mismatch. Those problems range from security aspects, maintainability, portability through 'simple' syntax and data-type mismatch. The market soon after witnessed the development of object-relational mapping (ORM) systems, which greatly reduced the amount of problems the programmers had to face [7,8].

ORM tools allow the programmer to focus on the code development without the need for advanced knowledge of SQL nuances. Although from the programmers point of view ORM systems are quite easy to use, the data transformations required in such mappings can be very complex, especially since they often involve advanced joins, nested queries and support for data update operations.

Nowadays object-relational mapping tools are available for most of the popular programming languages. Due to the popularity of their base languages the most noteworthy are Hibernate and JDO for Java, LinQ and ADO .NET for .NET platform, Django-models and SQLObject for Python. Attempts have been made by various groups to create a standard for the object relational mapping frameworks. One of the proposals for such a standard for Java language is the Oracle's Java Persistence API (JPA) [10]. It quickly gained popularity among software developers and nowadays there are many application servers implementing JPA. This includes JBoss, Apache Geronimo or Oracle's OC4J.

The authors have chosen to present the effects of their research on two ORM systems: Hibernate for Java [9,12,13], which is currently the most popular ORM system, and Django-models - a powerfull ORM for Python and a part of a very popular Django framework.

The initial research on joining ORM systems and a recursive queries technique has been made for a SQLObject - an ORM system for Python language, and a PostgreSQL DBMS [11]. Initial work on recursive query system for Hibernate has been presented in [6]. The following work extends both papers by, inter alia, presenting a complete approach to such extensions with efficiency tests

2 Contribution

The algorithms for generating recursive queries and its implementation for IBM DB2, Oracle and PostgreSQL database management systems are the main contribution of the following work. An algorithm for generating recursive query for PostgreSQL widely extends the algorithm presented in [11]. The difference between them will be presented in *Features* section. Because of the differences between SQL dialects supported by mentioned database systems and the SQL:99 standard the authors had to carefully analyze different scenarios and establish standardized query generation methodologies.

The second equally important result is the development and implementation of interfaces for recursive query definitions for Django-models and Hibernate ORM. These interfaces have been designed to retain maximum compatibility with standard mapping interfaces. In Hibernate the key issue was to design them according to the JPA standard convention. JPA standard and its implementation in Hibernate allows for two methodologies of mapping. The first one uses annotations for persistent objects' classes. The second one is based on XML documents storing configuration data. Therefore, the authors of the following work have developed separate interfaces for both of these methods.

The results of efficiency tests performed on the developed interfaces have been described in the *Performance* section. They show that querying recursive data using proposed enhancements to both Hibernate and Django-models can be accomplished in acceptable time. The native solutions gather the same data in a much longer time. The proposed solution is 20 times faster with 900 records data set up to 100 times faster with 4500 records data set. In that section the authors also present an explanation to these results.

3 Data Structures

The research presented in this papers focuses on the problem of processing graph and hierarchical data structures. There is a lot of real-life problems associated with such structures among which are: finding the communications links between the two cities or finding routes based on information provided by GPS systems, processing championships' scoreboards, corporate hierarchy or bill-of-material.

The following section presents two natural examples of recursive data. Example 1 present data concerning corporate hierarchy with special focus on employees hierarchy. Example 2 describes a network of flight connections between cites.

The example data are presented tables 1 and 2

Table 1. Hierarchical data, table Emp

empId	bossId	sname	fname	salary
7521	7698	Christie	Andrew	210
7566	7839	Jones	Brandon	360
7654	7698	Ford	Carl	210
7698	7839	Blake	Ernest	360
7782	7839	Bell	Gordon	360
7788	7839	Willis	James	360
7839		Smith	John	500
7844	7698	Turner	Johnathan	210
7902	7698	Adams	Trevor	210
7900	7566	Miller	Kyle	150

Table 2. Graph data, table Conns

departure	arrival	flightId	price	travelTime
Phoenix	Huston	PW 230	100	3h 10min
Huston	Chicago	RW 121	90	2h 45min
Huston	Dallas	RW 122	80	3h
Dallas	Chicago	DW 80	110	2h 30min
Chicago	Atlanta	CH 542	220	2h 45min
Chicago	Berlin	CH 543	360	7h 15min
Paris	Berlin	TW 118	300	1h 10min
Dallas	Berlin	DW 90	350	5h 45 min
Berlin	Boston	YW 421	100	6h
Chicago	Boston	CH 544	250	2h 15min

As a reminder, and a place for reference, let us have a look at a simple recursive CTE, that could be used for querying employees data structure (1.1):

Listing 1.1. List of Smith's Subordinates

```
WITH RECURSIVE rcte (
  SELECT sname, fname, empId, False as isSub
    FROM Emp WHERE sname = 'Smith'
 UNION
  SELECT e.sname, e.fname, e.empId, True as isSub
    FROM Emp e JOIN rcte r ON (e.bossId = r.empId)
)
SELECT sname, fname
  FROM rcte WHERE rcte.isSub = True;
```

4 Data Representation in Object-Relational Mapping

Object-relational mapping systems are the result of intersection of two worlds - object-oriented programming languages and relational databases. They play a major role in the development of modern database software. For most programmers it is not a question 'whether to use an ORM system', but rather 'which one'. There are many kinds of ORM tools currently available on the market. They differ among themselves as to the programming language they are designed for and the scope of supported databases management systems. Some of them are available as commercial products, while others have a fully open source code. One of the most popular ORM frameworks is Hibernate for Java.

Hibernate supports most of the major relational DBMSs. In accordance with JPA standard it offers two methods of defining an object-relational mapping. To map Java classes to database tables, developer may choose to define a mapping configuration in an XML document or to define it using Java Annotations. Most of the programmers decide to use XML files choosing automatic generation of the corresponding Java classes performed by Hibernate. Hibernate supports

mechanisms for automatic handling of one-to-one, one-to-many and many-to-many relation types. Sample objects generated using Hibernate framework may resemble the classes form listings 1.2.

Listing 1.2. Emp and Connections class representation

```
public class Emp {
  public long empId;
  public long bossId;
  public String sname;
  public String fname;
  public long salary;
  ...
```

```
public class Connections {
  public String departure;
  public String arrival;
  public String flightId;
  public String price;
  public double travelTime;
  ...
```

Django is a high-level Python framework used for rapid development of dynamic and powerfull information web-portals and applications. This tool has many features valued by web-developers including high scalability, efficiency in processing huge amounts of data requests from application users. It is equipped in a high-level ORM system known as django-models [5]. The corresponding classes to the above examples in Hibernate, when written in Django take the form presented in listing 1.3.

Listing 1.3. Python classes for Employees and Connections representation

```
from django.db import models

class Empl(models.Model):
    last_name = models.CharField(max_length=200)
    chief_id = models.IntegerField()

class Connections(models.Model):
    city_start  = models.CharField(max_length=200)
    city_end    = models.CharField(max_length=200)
    travel_time = models.FloatField()
```

Assuming that both data collections contain hierarchic data we need to process, the built-in 'native' solutions allow for only two methods for achieving this goal: sending a pure and complete SQL query to the server, and writing a loop in the host language that recursively sends queries to the DBMS at each step of the recursion. However, the evaluation time and network traffic in the second case are not acceptable. As for the first option, well, the whole idea behind ORM systems is not to use pure SQL (the reasoning behind this approach is well-known and falls outside of the scope of the following work).

The problem with long evaluation times comes from the fact that for each returned object, database server has to check if there are object with "subordinate" relation to that object. As a result, this query is being sent to the database as many times as there are employees in a sought structure. The solution proposed in this paper tested on a sample of 900 employees completed the evaluation process more than 20-times faster. The observed increase in evaluation speed has been achieved due to the usage of SQL's query type called recursive common

table expression or in short recursive query. Performed tests will be discussed in detail in *Performance* section.

5 Recursive Query in Hibenate

The authors have developed recursive query generators with automatic mapping of results onto objects. The author also provide interfaces for these generators. Each interface has been developed for handling both cases - configuration supplied through XML files and through annotations. The recursive extension to Hibernate's XML configuration has been presented in [6]. However, let us remind the structure of the mapping configuration proposed there. Listing 1.4 presents a configuration defined in XML document which allows for recursive querying of employees hierarchy.

Listing 1.4. Mapping configuration for travel connections

```
<rcte>
   <rcteTable name="travel" max-level="4" cycle="false"/>
   <tables>
     <table>Connections</table>
   </tables>
   <recursive-condition>
      <on>Connections.departure</on>
      <to>Connections.arrival</to>
   </recursive-condition>
   <summands>
     <sum>Connections.travelTime</sum>
     <conc>Connections.flightId</conc>
     <conc using=";">Connections.arrival</conc>
   </summands>
   <constants>
     <const>Connections.arrival</const>
   </constants>
   <filter section="seed">
      Connections.departure = $Param(departure)
   </filter>
   <filter section="outer">
    Connections.arrival = $Param(arrival)
   </filter>
</rcte>
```

Mapping through XML files has some notable advantages - it may be updated without the need for code recompilation. However, Hibernate's second method uses annotations of persistent object, which seems to be more elegant. For developers that prefer annotations rather than XML configuration the authors have prepared another set of generators. This solution is based on Java annotation attached to classes representing the results of required recursive query. Listing 1.5

presents a sample usage of annotations to specify recursive query traversing corporate hierarchy. An annotated class that would represent recursive search for flight connections would have a corresponding structure.

Listing 1.5. Emp.java file with annotations

```
package sample.recursive.mapping;
import org.ncu.hibernate.annotations.*;
@RecursiveQuery (maxLevel = 4)
@Tables (name = "Emp")
@RecursiveCondition (on= "Emp.bossID", to= "Emp.empId")
@Summands (conc = { "Emp.empId", "Emp.sname" })
@Filter (seed = "Emp.sname = $Param(sname)")
public class Suboridnates {
    @Column(name = "Emp.empID")
    public String id;
    ...}
```

Configuration of a given query should specify the tables used to collect data. It also should comprise the joining predicates for those tables and, in particular, recursion predicate. Besides those information the construction the authors propose includes parameters helpful in specifying additional options. Listing 1.4 presents an XML configuration document for travel connections problem,which contains most of the parameter nodes available for the programmer. The names of the XML nodes correspond to names of annotations that serve the same purpose. The authors have identified each syntax element occuring in recursive CTE and created a corresponding configuration element.

The root of the XML configuration document is called `rcte`. Its first child node is used to specify the name of the output RCTE table (lines 2 and 3 of provided listing). This node is called `rcteTable` and has up to three attributes: `name` used to define the actual name of the RCTE, `max-level` specifying recursion depth and an optional `cycle` attribute used to enable cycle protection if it is provided by DBMS vendor. Because in annotation system a programmer defines the name of the defined class explicitly, the annotation `RecursiveQuery` takes only `maxLevel` and `cycle` properties. The next node of the XML configuration document is called `tables`. Its child nodes called `table` are used to specify which database tables will be used to generate the result. Names of those tables are passed as character elements. In the presented example the only table needed to generate the result is Connections. It is represented by `<table>Connections</table>` code. The same role for annotation system play the `Tables` annotation and its `name` parameter.

Besides defining source tables, the programmer should supply information which table columns will be used in a recursion predicate. This information is stored in the `recursive-condition` node. This node has two child nodes: `on` and `to` which use character data to store corresponding column names. In the provided listing those nodes are placed in lines 7-10. The corresponding annotation is `RecursiveCondition`. Here the programmer also has to define the `on` and `to` columns as parameter values.

In addition to the required nodes describing source tables, developer may choose to specify additional options. Those include fields used to collect data such as sums or concatenations. For XML configuration, the main node for such information is `summands`. It allows for specification of unlimited amount of two types of child nodes: `sum` and `conc`. The `conc` node is used to specify column which will be used to create a concatenated string. Concatenated values are separated by default using single white space. However, a programmer may choose to specify optional attribute `using` of the `conc` node to overwrite this character with a chosen string. Corresponding annotation is `Summands` with parameters `conc` and `sum`. Besides collection fields a programmer may also specify a constant field if needed. An example of such definition is `<const>Connections.arrival</const>` (line 17 of the listing 1.4). This field may be used by the presented generator to optimize resulting query using predicate push in technique described in [4].

To specify additional filtering predicates a programmer may use `filter` nodes with `section` attribute or `Filter` annotation. Depending on the target subquery this configuration element may have three values: *seed*, *recursive* and *outer*. The programmer provides target predicates as character elements using *$Param()* function that enables passing of parameters in the Java source code.

Another XML node supported by the generator is the `outer` node. Its contents define additional properties of the outer SQL query that uses the rcte. This node may additionally contain `property` tags, which have the same meaning as identical elements in classical Hibernate configuration files. For example, they may be used to select only relevant columns instead of all generated ones. Similarly, the provided annotation query generator also allows for standard annotations used in Hibernate without putting on them any limitations.

In the paper [2] authors have shown that the performance of recursive queries in DB2 database system highly depends on existing indexes placed on the fields from recurrence predicate. Thus the authors of this paper have decided to add a special attribute `indexed = True/False` in the recurrence condition definition. If it is set to `True` the configurator checks if the proper fields are indexed and if not - inserts missing indexes into the database. With accordance to [2] this attribute is ignored by configurator for database systems other than DB2.

At first glance, this configuration may seem very complex. However there are clearly distinguished sections corresponding to the elements described. Names of the configuration nodes have been chosen so that they would be self-explaining in most cases. Based on such configuration the query generator will construct a corresponding recursive SQL query. The resulting query compared with the configuration shows the transparency of the latter. The usage of newly generated objects for both methods is exactly the same as standard objects generated by Hibernate thus it will not be discussed here.

The algorithms from previous papers in this topic presented only limited capabilities. The initial paper [11] presented an algorithm that worked only with PostgreSQL and only with one source table. Also, it had no filter clauses, summands, concats etc. The second attempt aimed solely for Hibernate's XML configuration, and it did not fully comply with the JPA standard. It also lacked

the standardized approach to cycle-protection and did not implement predicate push-in optimization. The algorithms presented in this paper fulfill these features, thus they are much more useful than the previous one.

6 Recursive Query in Django-Models

Django-models do not provide easy mechanism for recursive data processing. The closest solution is the same as for Hibernate - some form of a loop that sends huge amounts of queries to a DBMS. Programmers have to face the problems of long execution times and large amounts of data transfer - often useless data. Having already identified the elements of the RCTE, we will present how they could be incorporated in another ORM system.

In django-models the parameters (corresponding with Hibernate solution) are passed as variables in an object inheriting from the object `Recursive`. In this solution we have included an additional parameter: `original`. It allows a programmer to restrict the column list from the outer SELECT query to the columns which names correspond to the source table's column names. It gives the programmer more freedom in restricting the size of the result table.

Listing 1.6. Example of a recursive django-model object

```
from test_app.models import Connections
from recursive import Recursive

class RecTravel(Recursive):
    echo = True
    table = Connections
    recursiveOn = 'city_start'
    recursiveTo = 'city_end'
    maxLevel = 4
    summands = ["travel_time"]
    concats = ["city_end", "id"]
    outerFilters = ["city_end='Warsaw'"]
    seedFilters = ["city_start='Toronto'"]
    original = True
    cycle = False
```

The parameter `echo` allows the user to view the SQL query generated out of the given class definition. For the example 1.6, the resulting query would have the same form as the query generated by hibernate in the example 1.4.

7 Performance

This section presents some of the tests that compare native ORM methods and the proposed extension for both Hibernate and Django. The problem of corporate hierarchy has been tested in five cases: 900 records with 7 levels of hierarchy,

1800 records with 8 levels of hierarchy, 2700 records with 9 hierarchy levels, 3600 records with 10 hierarchy levels and 4500 records with 11 hierarchy levels. To test the native Hibernate's method, the authors prepared a source code based on the *while* loop.

A simple analysis of this code's composition reveals that during the execution, Hibernate generates as many queries and their calls as there are matching objects in the requested structure. In comparison, the recursive CTE query sent to the DBMS is calculated in the time similar to a union of as many select queries as the depth of the recursion increased by one. The average results of those tests are presented in tables 3, 4 and 5. Tests were performed on different machine configurations for different database systems[1]. The point of the tests was to compare the evaluation of Hibernate's solutions and the proposed extension. The *Hibernate loop* column presents the time needed to complete the execution of the "native" Hibernate code. The *ratio* column presents the percentage of time needed for the RCTE extension to complete in comparison to the native method. The figure 1 presents those results on a chart.

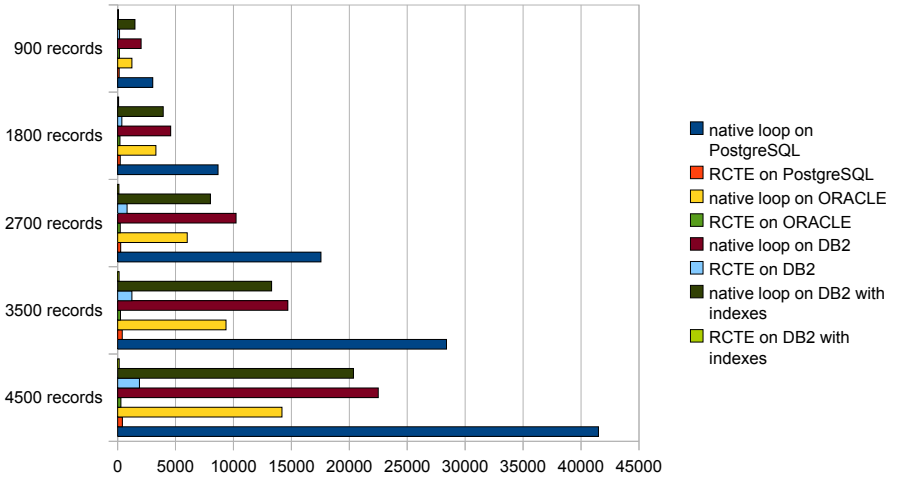


Fig. 1. Performence tests on PostgreSQL

The table 4 shows the performance of recursive querying without index on the `boss_id` field and with them. Both test cases have been executed on the same machine configuration.

For django-models the tests have been based on exactly the same databases as in previous tests. The recursive query was created out of `recEmpl` class which contained identical parameters as Hibernate's test class. The native django solution has been constructed using the `for` loop.

[1] Thus they say nothing about comparison between database system performance.

Table 3. Comparison of average execution times on ORACLE

	Hibernate loop	Hibernate RCTE	RATIO
900 rec.	1237 ms	162 ms	**13.1 %**
1800 rec.	3303 ms	201 ms	**6.08 %**
2700 rec.	6012 ms	232 ms	**3.85 %**
3500 rec.	9354 ms	256 ms	**2.74 %**
4500 rec.	14192 ms	289 ms	**2.04** %

Table 4. Comparison of average execution times on IBM DB2

	without indexes			with indexes on `boss_id`		
	loop	RCTE	RATIO	loop	RCTE	RATIO
900 rec.	2026 ms	171 ms	8.44 %	1498 ms	78 ms	**5.2 %**
1800 rec.	4584 ms	371 ms	8.09 %	3936 ms	86 ms	**2.18 %**
2700 rec.	10225 ms	820 ms	8.02 %	8011 ms	111 ms	**1.38 %**
3500 rec.	14698 ms	1233 ms	8.39 %	13291 ms	135 ms	**1.01 %**
4500 rec.	22495 ms	1887 ms	8.39 %	20356 ms	139 ms	**0.68 %**

Table 5. Comparison of average execution times on PostgreSQL

	Hibernate loop	Hibernate RCTE	RATIO
900 rec.	3033 ms	143 ms	**4.71 %**
1800 rec.	8669 ms	266 ms	**2.61 %**
2700 rec.	17550 ms	271 ms	**1.54 %**
3500 rec.	28391 ms	405 ms	**1.43 %**
4500 rec.	41500 ms	414 ms	**1.00 %**

Table 6. Results of efficiency tests for django-models solution

Amount of records	**Django RCTE**	**Native Django code**	**Ratio**
900	84 ms	2024 ms	**4.15 %**
1800	165 ms	4223 ms	**3.91 %**
2700	247 ms	7571 ms	**3.26 %**
3600	318 ms	10885 ms	**2.92 %**
4500	413 ms	19612 ms	**2.11 %**

8 Conclusions and Future Work

In conclusion, the results of the tests show that the solution proposed by the authors allows for benefiting from the advantages of Object Relational Mapping in applications using recursive data. As it has been described both presented configuration methods preserve the original Hibernate's configuration style which makes them easily accessible for the user.

Next the authors plan to continue the research on providing support for another features implemented in database management systems at the level of Object Relational Mapping. One of the research subjects is support for logical constraints that would allow for automatic generation of DBMS specific triggers. Another challenge would be to integrate the support for dynamic SQL into an ORM framework. Parallel to that research the authors plan to develop corresponding extensions for other Database Management Systems with the special focus on SQL Server and SQL Anywhere.

References

1. Brandon, D.: Recursive database structures. J. Comput. Small Coll. 21(2), 295–304 (2005)
2. Przymus, P., Boniewicz, A., Burzańska, M., Stencel, K.: Recursive Query Facilities in Relational Databases: A Survey. In: Zhang, Y., Cuzzocrea, A., Ma, J., Chung, K.-i., Arslan, T., Song, X. (eds.) DTA and BSBT 2010. Communications in Computer and Information Science, vol. 118, pp. 89–99. Springer, Heidelberg (2010)
3. 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)
4. Burzańska, M., Stencel, K., Wiśniewski, 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)
5. django-models, https://docs.djangoproject.com/en/dev/topics/db/models/
6. Burzańska, M., Boniewicz, A., Szumowska, A., Wiśniewski, P.: Hibernate the Recursive Queries - Defining the Recursive Queries Using Hibernate ORM. To appear in Proceedings of ADBIS 2011 (2011)
7. Melnik, S., Adya, A., Bernstein, P.A.: Compiling mappings to bridge applications and databases. ACM Transactions on Database Systems (TODS) 33(4), 1–50 (2008)
8. Keller, W.: Mapping objects to tables: A pattern language. In: EuroPLoP (2007)
9. Hibernate, http://www.hibernate.org
10. DeMichiel, L.: Java Specification Requests JSR 317: JavaTM Persistence 2.0 (2009), http://jcp.org/en/jsr/detail?id=317
11. Burzańska, M., Stencel, K., Suchomska, P., Szumowska, A., Wiśniewski, P.: Recursive Queries Using Object Relational Mapping. In: Kim, T.-h., Lee, Y.-h., Kang, B.-H., Ślęzak, D. (eds.) FGIT 2010. LNCS, vol. 6485, pp. 42–50. Springer, Heidelberg (2010)
12. Bauer, C., King, G.: Java Persistence with Hibernate. Manning Publications Co., Greenwich (2006)
13. O'Neil, E.J.: Object/relational mapping 2008: hibernate and the entity data model (edm). In: Proc. ACM SIGMOD, pp. 1351–1356 (2008)

Partial Aggregation Using Hibernate

Michał Gawarkiewicz and Piotr Wiśniewski

Faculty of Mathematics and Computer Science
Nicolaus Copernicus University
Toruń, Poland
{garfi,pikonrad}@mat.umk.pl

Abstract. In this paper authors describe tools allowing for the use of partial aggregation techniques in analytical queries from the level of object-relational Hibernate mapping. Grouping queries are quite burdensome for databases. Therefore, in applications processing large amounts of data, partial aggregation is made (it should not be confused with the huge data sets, which are aggregated in data warehouses). Unfortunately, applications written using object-relational mappings cannot use this type of solutions directly. The authors decided to fill this void by adding easy-to-use tools to automatically create aggregations and very simple mechanisms to retrieve the aggregated data.

1 Introduction

The use of database applications in business additionaly to making work easier causes a positive side effect of collecting large amounts of data. These data allows carrying out a series of studies, etc. These analyses require grouping queries that are much heavier to execute. In the case of huge data sets, this aspect led to the development of a large branch of science - data warehouses. This work applies to smaller collections, for which the application of data warehouse appears to be pointless, but large enough, so that direct analytic queries would have unacceptable execution times.

In practical applications this problem is easy to get around by preparing additional aggregation tables and setting triggers on the appropriate events in the database. Then in the queries, rather than refering to the data directly, analytic queries use partially aggregated data.

Parallel to the development of relational databases, the field of object-oriented programming languages and data modeling methods evolved as well. Modeling of relational data structures resembles more and more the modeling process of objects and classes. The issue has inspired the research on techniques of mapping objects to relations and relations to objects [1,2]. The data transformations required in such mappings can be very complex, especially since they often involve advanced joins, nested queries and support for data update operations.

Object Relational Mapping tools allow the programmer to focus on the code development without the need for advanced knowledge of SQL nuances. This

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 90–99, 2011.

results in reduction of the time needed to develop software and an increased readability of the source code. Additional benefit of using object-relational mapping tools is the increased source code maintainability and portability between different DBMSs. The number of supported DBMSs depends on the actual ORM framework used - from one particular vendor to a full support of all major DBMSs.

Nowadays object-relational mapping tools are available for most of the popular programming languages. Due to the popularity of their base languages the most noteworthy are Hibernate and JDO for Java [3], LinQ and ADO .NET for .NET platform and SQLObject for Python. Because nowadays Java is the most popular language for software development we will restrict our discussion to it.

Unfortunately, application designers using object-relational mapping can not directly use the technique of partial aggregation. To use it they must somehow manually create the appropriate tables and triggers on tables with raw data. As a result, the clarity of solutions is lost and subsequent development and maintenance of such a code becomes more expensive. These facts led the authors to attempt to integrate the mentioned techniques with object-relational mappings, and the results of this research are presented in the following paper.

The presented work is another after [4] and [5] of a series of works complementing the object-relational mapping's support for advanced techniques for modern object-relational systems.

2 Contribution

In this paper, automatic algorithms have been developed to define auxiliary tables for storing partially aggregated data and to generate triggers that fill these tables. In addition, an easy to use interface using annotation to use these algorithms, maintaining JPA [6] standards was designed. The third aspect is the simple built-in HQL language parser that recognizes whether the sent request is supported by the stored aggregations and, if so, takes over its execution.

3 Motivation Examples

Let us consider sales data stored in the schema presented on figure 1.

An example of an analytical question is the amount of each product sold in the selected time interval. In the case of small amounts of data, we can accomplish the query as follows:

Example 1.

```
SELECT invoiceLines.product_id, invoice.date,
       sum(invoiceLines.quantity)
  FROM invoice JOIN invoiceLines USING (invoice_id)
 GROUP BY invoice.date, invoiceLines.product_id
HAVING date BETWEEN '2011-07-16' AND '2011-07-22'
```

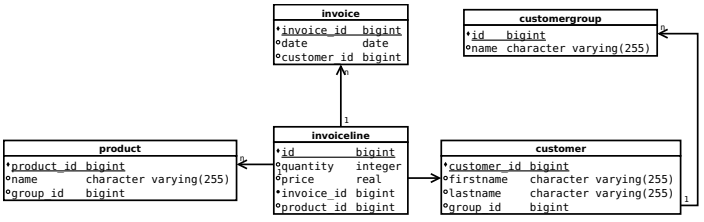


Fig. 1. Base Data Schema

In the case of huge amounts of data, such an approach is highly inefficient, which resulted in the development of data warehouses. However, there is a wide range of large sets, in which the use of warehouse-type solutions has no economic justification, while the use of direct questions is unacceptable at a time. The solution is then the partial aggregation of data. The above scheme should be extended to `dw_invoice_date_product` table presented on the figure 2:

dw_invoice_date_product	
id	serial
sum_quantity	real
date	date
product_id	bigint

Fig. 2. dw_invoice table

Then triggers on insert, update and delete should be applied on events modifying data in the invoiceLine table.

Then instead of the above query, the following could be used:

```
SELECT product_id, date, sum_quantity
 FROM dw_invoice_date_product
 WHERE function = 'sum'
       AND date BETWEEN '2011-07-16' AND '2011-07-22'
```

Of course, time to complete this query is incomparably shorter than the execution time of the query from the previous example.

The second example is the sum of sales for selected groups of customers, and monthly sale totals for each customer. The relevant queries are as follows:

Example 2.
Sum of sales for selected groups of customers:

```
SELECT customer.group_id, invoice.date,
       sum(invoiceLine.quantity)
 FROM customer JOIN invoice Using (customer_id) JOIN
```

```
      invoiceLine USING (invoice_id)
 GROUP BY customer.group_id, invoice.date
 HAVING date BETWEEN '2011-07-16' AND '2011-07-22'
```

Monthly sale totals for each customer:

```
SELECT invoice.customer_id, month(invoice.date),
       year(invoice.date),
       sum(invoiceLine.quantity * invoiceLine.price)
 FROM invoice JOIN invoiceLine USING (invoice_id)
 GROUP BY invoice.customer_id, month(invoice.date),
          year(invoice.date)
```

To deal with these aggregations, one can add a table aggregating sales information by customers and date, and then from it derive separate aggregations for client groups and monthly sales, and place them in two additional tables. As a result, the scheme should be extended to the structure shown in the figure 3.

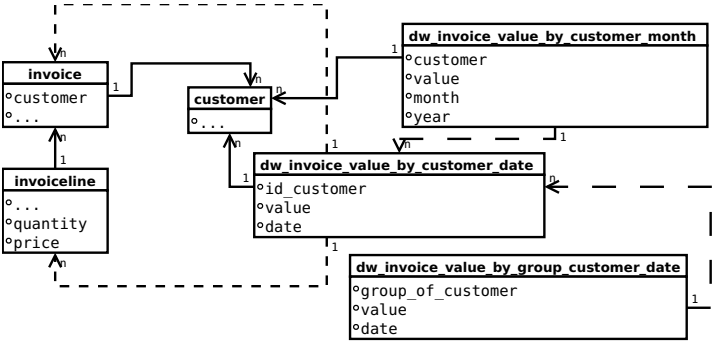


Fig. 3. Aggregation tables

4 Hibernate

Object-relational mapping systems are the result of intersection of two worlds - object-oriented programming languages and relational databases. They allow for separation of the business logic layer from database layer and as a result - increased portability and maintainability of software's source code. Additional feature of ORM frameworks which has a big impact on their popularity is the speed of source code developments during which only minimal knowledge is needed about the advanced aspects of SQL language. There are many kinds of object relational mapping tools currently available on the market. They differ among themselves as to the programming language they are designed for and the

scope of supported databases management systems. Some of them are available as commercial products, while others have a fully open source code. One of the most popular ORM frameworks is Hibernate for Java language.

Hibernate supports most of the major relational DBMSs. In accordance with JPA [6,7] standard it offers two methods of defining an object-relational mapping. To map Java classes to database tables, developer may choose to define a mapping configuration in an XML document or to define it using Java Annotations with nice Entity data model [8]. Sample objects generated using Hibernate framework may resemble the classes form listings 1.1, 1.2, 1.3.

Listing 1.1. Invoice class representation

```
@Entity
public class Invoice {
    private Long id;
    private Date date;
    private Customer customer;
    private List<InvoiceLine> invoiceLines;
    ...
```

Listing 1.2. Customer class representation

```
@Entity
public class Customer {
    private List<Invoice> invoices;
    private Long id;
    private String firstName;
    private String lastName;
    private CustomerGroup customerGroup;
    ...
```

Listing 1.3. CustomerGroup class representation

```
@Entity
public class CustomerGroup {
    private List<Customer> customers;
    private Long id;
    private String name;
    ...
```

From the Hibernate, queries from the examples 1 and 2 can be expressed as follows:

Example 3. Amount of each product sold in the selected time interval:

```
SELECT invoiceLines.product_id, invoice.date,
       sum(invoiceLines.quantity)
  FROM Invoice invoice JOIN invoice.invoiceLines invoiceLines
 GROUP BY invoice.date, invoiceLines.product_id
HAVING date BETWEEN '2011-07-16' AND '2011-07-22'
```

Sum of sales for selected groups of customers:

```
SELECT customer.customer_group, invoice.date,
       sum(invoiceLines.quantity)
 FROM Invoice invoice JOIN invoice.customer customer JOIN
      invoice.invoiceLines invoiceLines
 GROUP BY customer.customer_group, invoice.date
 HAVING date BETWEEN '2011-07-16' AND '2011-07-22'
```

Monthly sale totals for each customer:

```
SELECT invoice.customer_id, month(invoice.date),
       year(invoice.date),
       sum(invoiceLines.quantity * invoiceLines.price)
 FROM Invoice invoice JOIN
      invoice.invoiceLines invoiceLines
 GROUP BY invoice.customer_id, month(invoice.date),
          year(invoice.date)
```

Grouping queries sent to the database will be of course similar to the queries presented in the examples 1, 2 - the differences are cosmetic resulting from the automatic code generation. Hence, they are executed with an appropriate big time consumption for large data sets. Attempts to use the solutions described in the previous section, so far required the designer to give up elegance of Hibernate and manually create additional tables and write by hand triggers that fill these tables. Gathering of information can be carried out semi-automatically - these tables can be viewed as non-modifiable objects on the Hibernate side. Although this approach solves the aspect of the execution time, it significantly increases the cost of producing and maintaining such a written software. Any modification of the original scheme requires an analysis of the impact on the additional tables, rebuilding the extended scheme and rewriting the triggers. These problems led the authors to try to automate these issues.

5 Proposed Solution

The presented solution involves several aspects. The first is a set of annotations, through which it is defined which aggregations on the given classes are interesting. The generator implemented by the authors analyses these annotations to generate tables for storing aggregations and creates triggers filling them with data. The next aspect is the aggregated data retrieval interface to parse the query in HQL language, and when it detects support for supported aggregation, it takes information from the aggregated data. In case where there is no support, the HQL query is executed by native Hibernate engine.

Annotations

The proposed solution introduces annotations beginning with the prefix @DW.. Aggregated fields get annotation @DWAgr(funtcion="...") and fields where a

grouping is to be held get annotation @DWDim. The following example illustrates the use of them in the first example:

Listing 1.4. InvoiceLine class representation

```
@Entity
public class InvoiceLine {
    ...
    @DWDim(Dim = "date")
    private Invoice invoice;
    private Long id;
    @DWDim
    private Product product;
    @DWAgr(function="SUM")
    private Integer quantity;
    ...
```

Dim extension in the annotation @DWDim(Dim = "date") indicates that the grouping is done with granularity of invoice dates, rather than invoices alone.

In the above example the generator implemented by the authors will create a table `dw_invoiceline_by_product_date` with fields `id_product`, `date`, `sum_quantity` and triggers on events on the invoiceLine table that will care about data relevance in the aggregate table.

Function parameter in the DWAgr annotation specifies which aggregate functions will be used. At this stage of our solution, basic aggregate functions are supported, ie. SUM, MIN, MAX, AVG and COUNT. In the case of simple values, relevant values of cells in tables are subject to aggregation. The case of the compound values will be discussed after the next example.

In the second case when update affects aggregation for a customer or customer group, annotation mechanism is pinned to the invoice class:

Listing 1.5. Invoice class representation

```
@Entity
public class Invoice {
    private Long id;
    @DWDim(Extention="month()")
    private Date date;
    @DWDim(Extention="group")
    private Customer customer;
    @DWAgr(function="SUM(quantity*price)")
    private List<InvoiceLine> invoiceLines;
    ...
```

The above example introduces several enhancements compared with the previous one. First dw_invoice_by_date_customer table is built and then triggers on events on invoice and invoiceLine tables. They will fill aggregations in those tables. In this example we need triggers on both tables, since modifications on both header data and a single record in invoiceLine table affect aggregations.

For the presented solutions it is not a problem that aggregate function is set on the field being a list, which means that the data are taken from a separate table. Hibernate metadata analysis is sufficient to create a properly working triggers by the implemented automatic generator. It is important that the relation is many invoiceline records into one invoice record. This way granulation of grouping described at the invoice level clearly distributes which records with aggregations get relevant data from invoiceline. InvoiceLine field contains a list of objects. This list relates to the fields of an other table. In this situation aggregation can apply not only to specific columns, but also as presented in the case to algebraic expressions on columns. In the current version only simple algebraic expressions without control instructions and nested parentheses is supported.

The described case in the second example expects aggregation at further level, ie collecting customer sale totals for each month. Parameter (Extension=month()) of DWDim annotation on the date field informs the generator about the need to create a `dw_invoice_by_month_date_customer` table that will aggregate data from the `dw_invoice_by_date_customer` table. This aggregation will be held through triggers generated for the `dw_invoice_by_date_customer` table. Parenthesis in this parameter tell the generator that it is a SQL function calculated from the value of the field, where a date field has a special procedure for the analysis of functions. The indication of the month will automatically add the year column in which the year will be saved, so that the aggregations from various years and the same month will not get together. If a programmer wants such a distinction, he or she should write @DWDim(Extention="month() only").

The second use of Extension parameter concerns customer field, which is a pointer to a complex object and is represented on the database side by the customer table tuples. This extension indicates generator the need to establish `dw_invoice_by_date_group_customer` table in which aggregations are collected for the client sums. As before, this is done by triggers created on the `dw_invoice_by_date_customer` table.

Getting Aggregated Data

The presented extension of Hibernate introduces two methods of access to aggregated data. The first one allows for data retrieval according to the QBE model. ncu.Agregations class has a static method qbe. As a result, objects of aggregation are returned. These features will be presented on the example of usage based on the first example:

```
// date1 is set to '2011-07-16', date2 is '2011-07-22'
ncu.Agregations.qbe(InvoiceLine.class, "SUM(quantity)",
  new Object[] {Date.class, Product.class},
  new Object[] {ncu.Agregations.helpers.between(date1, date2)});
```

This call executes query:

```
SELECT SUM(sum_quantity), date, id_product
  FROM dw_invoiceline_by_product_date
  GROUP BY date, id_product
  HAVING date BETWEEN '2011-07-16' AND '2011-07-22'
```

This method has four parameters, the last one is optional. The first one specifies the class that aggregations relate to. This should be a class that contains definitions of aggregations discussed above. The second parameter specifies the aggregation functions we want to use. Of course, if the proposed solution could answer these questions, then these functions should be declared in the class definition in the @DWAgr(function...) annotation form. The third parameter defines the grouping as a list of objects. These can be classes or functions that were indicated in @DWDim annotation in the base class. The fourth parameter is a list of restrictions. In the example we used the helper function specifies that dates should be contained in a specified range.

This function call for the second example:

```
ncu.Agregations.qbe(Invoice.class, "SUM(quantity*price)",
  new Object[] {"month(date)", CustomerGroup.class},
  new Object[] {});
```

This call executes query:

```
SELECT SUM(value), date, group_of_customer
  FROM dw_invoice_value_by_group_customer_date
  GROUP BY date, group_of_customer
```

The second methodology uses a simple HQL parser implemented by the authors in `ncu.Aggregation` class as follows:

```
ncu.Agregations.hql("
SELECT invoiceLines.product_id, invoice.date,
       sum(invoiceLines.quantity)
  FROM Invoice invoice JOIN invoice.invoiceLines invoiceLines
 GROUP BY invoice.date, invoiceLines.product_id
HAVING date BETWEEN '2011-07-16' AND '2011-07-22'
");
```

The parser analyzes HQL query. If it recognize that query could be realized using prepared aggregation, the qbe method is called. If not than the original Hibernate HQL engine is called.

6 Future Work

In the future work the authors plan to extend the presented algorithms to work with ORACLE and IBM DB2 databases. It is also planned to aggregate more

statistical algorithms and make closer integration with the HQL language. A major weakness at this stage is the issue of inherited classes. Algorithms supporting aggregations on more complex base classes diagram containing inheritance will be an important challenge for further work.

References

1. Melnik, S., Adya, A., Bernstein, P.A.: Compiling mappings to bridge applications and databases. ACM Transactions on Database Systems (TODS) 33(4), 1–50 (2008)
2. Keller, W.: Mapping objects to tables: A pattern language. In: EuroPLoP (2007)
3. Hibernate, http://www.hibernate.org
4. Burzańska, M., Stencel, K., Suchomska, P., Szumowska, A., Wiśniewski, P.: Recursive Queries Using Object Relational Mapping. In: Kim, T.-h., Lee, Y.-h., Kang, B.-H., Ślęzak, D. (eds.) FGIT 2010. LNCS, vol. 6485, pp. 42–50. Springer, Heidelberg (2010)
5. Szumowska, A., Boniewicz, A., Burzańska, M., Wiśniewski, P.: Hibernate the Recursive Queries - Defining the Recursive Queries Using Hibernate ORM. In: ADBIS 2011 (2011)
6. DeMichiel, L.: Java Specification Requests JSR 317: JavaTM Persistence 2.0 (2009), http://jcp.org/en/jsr/detail?id=317
7. Bauer, C., King, G.: Java Persistence with Hibernate. Manning Publications Co., Greenwich (2006)
8. O'Neil, E.J.: Object/relational mapping 2008: hibernate and the entity data model (edm). In: Proc. ACM SIGMOD, pp. 1351–1356 (2008)

High Speed Optical Coherent Transmission System Using Narrowband FM Subcarrier Multiplexing

Hae Geun Kim

School of Computer and Information Communication,
Catholic University of Daegu,
330 Kumrak-ri, Hayang-up, Kyungsan-si, 712-702, Korea
kimhg@cu.ac.kr

Abstract. A high speed optical coherent transmission System using Narrowband Frequency Modulation (NBFM) subcarrier multiplexing (SCM) has been introduced where the FM modulated optical signal for MPSK digital data is transmitted by a single mode fiber. At the receiver, a photodiode (PD) generates the electrical signal including *N* channel SCM signals and the optical intermediate frequency (IF) from the coupler. The electrical signal from the BPF output is FM demodulated and then divided into the *N* SCM data.

We have analyzed the variance of Gaussian noise at the FM discriminator output for four adjacent sucarrier channels which are degraded by 10.1, 12.0, and 14.5 [nW/Hz] as the subcarrier frequency is increased by 0.5 GHz, respectively. The caculation results of BERs for the *M*-ary PSK modulators, BPSK, QPSK, 8-PSK, and 16-PSK schemes that shows the SNR differences between the modulation schemes are 1.9, 2.1. 2.3 dB at BER = 10^9, as the subcarrier frequency is increased by 0.5 GHz, respectively.

Keywords: Subcarrier Multiplexing, Narrowband Frequency Modulation, Optical Coherent Transmission.

1 Introduction

Recently, the transmission technologies such as TDM (Time Division Multiplexing), WDM (Wavelength Division Multiplexing), and the combination of those two methods are introduced to increase the optical bandwidth efficiency [1]. Optical TDM systems with data rates of more than 10 Gb/s suffer from chromatic dispersion. On the other hand, WDM is coping with electronic data limitation for the optical baseband transmission. Since WDM can accommodates both the analog and the digital signals [2], SCM and WDM can be an effective combination to transmit the high speed optical data.

Coherent subcarrier multiplexing (SCM) is also attractive choice for high speed optical data transmission, because it does not need the very high speed electronic devices. Thanks to the recent development of electronic devices, SCM can handle the

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 100–107, 2011.

high speed electronic signal. In a coherent SCM system, multichannel transmission can be performed in the radio frequency (RF) domain with one optical carrier. Here, the RF devices are more mature than the optical devices. Coherent optical fiber communication systems offer improved sensitivity over direct detection systems and multichannel transmission capability by using a tunable local laser [3][4].

In this paper, the NBFM has been used in the SCM system, since the signal utilizes the bandwidth similar to FDM (Frequency Division Multiplexing) and modulates the optical signal easily. Section 2 presents the conception of SCM coherent optical communications and the noise characteristics of the FM noise and its bandwidth, theoretically. Section 3 discusses the calculation of the variance of Gaussian noise at an FM discriminator output for the SCM system using *M*-ary modulation schemes. Then the resulting data are used in calculating BERs which compare with the theoretical performances of the modulators within the bandwidth. Section 4 discusses the summary and conclusions.

2 FM SCM Optical Coherent System

Fig. 1 depicts the transmitter and receiver of the SCM system using FM as a principal modulation scheme where *N* SCM with electrically modulated signals are summed and enter into an FM modulator. Then the laser diode signal at λ_i, where $i = 1,\ldots, k$, is FM modulated for transmission through the single mode fiber.

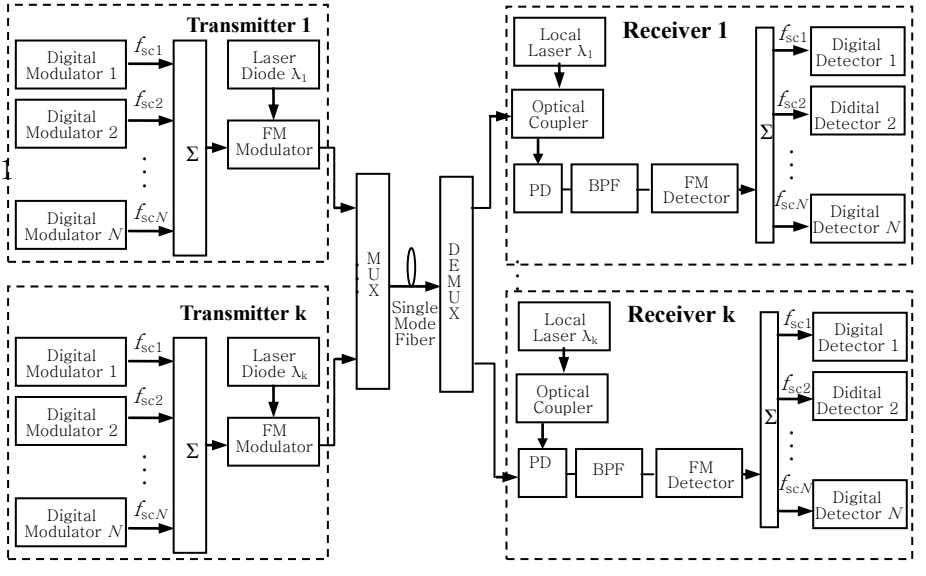


Fig. 1. *N* channel optical SCM system with a single mode fiber

At the optical receiver, the FM modulated signal is mixed with the local laser signal by an optical coupler, where we have used narrowband FM scheme throughout this paper. A photodiode (PD) generates the electrical signal including N channel SCM signals and the optical intermediate frequency (IF) signal which has the frequency different between two laser signals at λ_k. The electrical signal from the BPF output is FM demodulated and then divided into the N SCM data. The SCM for an MPSK signal can be expressed in an exponential form as:

$$Z(t) = A_{sc} e^{j\{\omega_{sc} t + \phi_m(t)\}} \tag{1}$$

where A_{sc} is an amplitude of subcarrier signal. ϕ_m is the phase shift, ω_{sc} is a radian value of a subcarrier frequency, and $m = 0, 1. \cdots, M\text{-}1$

$Z(t)$ can not be transmitted through real valued channel, so that be N channel SCM signal $X_N(t)$ can be presented the real value signal as:

$$X_N(t) = A_{sc}\,\mathrm{Re}\left[\sum_{i=1}^{N} e^{j\{\omega_{sc_i} t + \phi_{m_i}(t)\}}\right] = A_{sc}\sum_{i=1}^{N}\cos\{\omega_{sc_i} t + \phi_{m_i}(t)\} \tag{2}$$

where $\mathrm{Re}\{x\}$ denotes the real part of complex x. A laser diode output is modulated by $X_N(t)$. When we consider single transmitter, the outgoing optical signal at λ_i is

$$i_{FM}(t) = \sqrt{P_s}\cos\left\{\omega_s t + \omega_d \int_{-\infty}^{t}\sum_{i=1}^{N} A_{sc_i} X_N(\tau) d\tau\right\} \tag{3}$$

where P_S = Peak power of transmitter laser, $\omega_S = 2\pi f_s$, radian value of a transmitter laser frequency of, ω_d = Radian value of frequency deviation.

In a demodulator, a local laser signal can be expressed as

$$i_{LO} = \sqrt{P_{LO}}\cos\omega_{LO} t \tag{4}$$

where a local oscillator has a peak power P_{LO} with an angular frequency, ω_{LO}. When a local laser signal is mixed with i_{LO}, ω_{IF} with the intermediate frequency is generated. So, the BPS output can be

$$v_{IF}(t) = 2R\sqrt{P_{LO} P_S}\cos\{\omega_{IF} + \Phi(t)\} + n(t) \tag{5}$$

where R is the responsivity of photodiode, $\Phi(t)$ is an FM signal modulated by an MPSK signal, and $n(t)$ = white Gaussian noise.

Let white Gaussian noise be added at the transmission channel. The output noise of BPF in (5) can be written in the narrowband representation as:

$$n(t) = n_c(t)\cos\omega_{IF} t - n_s(t)\sin\omega_{IF} t = r_n(t)\cos\{\omega_{IF} t + \theta(t)\} \tag{6}$$

where n_C and n_S are the quadrature components of the noise, $r_n(t) = \sqrt{n_c^2(t) + n_s^2(t)}$ which has the Rayleigh-distributed noise envelope, and $\theta(t) = \tan^{-1}[n_s(t)/n_c(t)]$, which has uniformly distributed phase.

$$v_{IF}(t) = R(t)\cos[\omega_{IF}t + \psi(t)] \tag{7}$$

where $A_{IF} = \sqrt{P_{LO}P_S}$, $R(t) = \sqrt{[A_{IF} + n_c]^2 + [n_s(t)]^2}$, and $\psi(t) = \tan^{-1}[n_s(t)/(A_{IF} + n_c(t))]$

If the signal-to-noise ratio (SNR) is large (most of the time), $|n_C(t)| << A_L$ and $|n_S(t)| << A_L$. When the output signal of limiter passes through a discriminator with an envelop detector, the output of an envelope detector can be written as:

$$v_E(t) = \frac{A_L}{2\pi}\left[\omega_{IF} + \frac{1}{A_{IF}}\frac{d}{dt}n_s(t)\right] \tag{9}$$

A PSK detector is used in extracting the source data from discriminator output signal. For simplicity, if we consider single channel SCM signal with the noise, the baseband filter output can be written as:

$$v_B(t) = \frac{A_L\omega_d}{2\pi}\cos\left[\omega_{sc_i} + \phi_{m_i}(t)\right] + n(t) \tag{10}$$

where $n(t)$ is Gaussian noise. After the signal is passed through the baseband filter with cut-off frequency W, the noise autocorrelation given by (10) becomes

$$R_n(\tau) = \frac{A_n}{\pi}\int_0^W \omega^2\cos\omega\tau\, d\omega = \frac{2A_n}{\pi}\cos W\tau - \frac{4A_n}{\pi\tau^3}\sin W\tau + \frac{2A_nW^2}{4\pi\tau}\sin W\tau \tag{11}$$

We have calculated the the autocorrelation function using this equation for the NBFM case in the next section.

3 Calculation of the Noise Variance and Performance Evaluation

In order to investigate the problems that arise in science and technology, it is helpful to construct models which can approximate the real system. It is referred not only to tangible models but also to mathematical/computation models that give an approximated description of the behavior of a real system.

We calculated the variance of noise within the FM bandwidth of the SCM system. In Fig. 2, the spectrum of the narrowband FM bandwidth is illustrated, where the subcarrier signals with the FM quadratic noise which is filtered with a BPF is shown. The resulting output of the receiver is Gaussian noise with a quadratic form.

The calculation results of the autocorrelation function of (11) at the bandwidth of 0.05 ~ 0.9 GHz in nW/Hz for the subcarrier frequency at 0.1, 0.15, 0.2, and 0.25 GHz

are shown in Fig. 3. Here the noise variance is stable value at the bandwidth, W < 0.6 GHz for all subcarrier signals. In this range, the nose variances between adjacent sucarrier channels are 10.1, 12.0, and 14.5 [nW/Hz], respectively. So, the system performance is degraded as the subcarrier frequency increases.

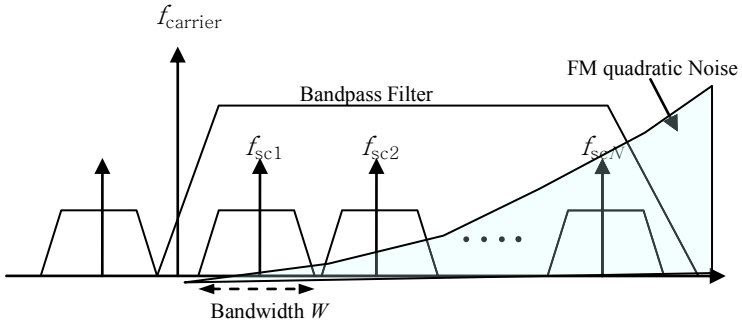


Fig. 2. Spectrum of the narrowband FM bandwidth of the proposed system detecting the subcarrier signals with the FM noise.

In Fig. 4, we calculated the signal-to-noise ratio (SNR) based on the calculation results in Fig. 3. In Fig.4, the SNRs for all subcarriers are also stable state at W < 0.6 GHz for all the subcarrier frequency. In this bandwidth range, the SNR difference of each subcarrier is less than 1 dB. In order to determine the SNR based on the calculated noise variance, the average energy of the two binary digits at the PSK detector output is A^2T where we let A, the amplitude of the carrier be 1. The output SNR can be calculated as T/σ^2 where T is 20 ns and σ^2 is the calculated noise variance at W = 0.6 GHz. So we have calculated the BERs of digital systems, BPSK, QPSK, 8-PSK, and 16- PSK, at the bandwidth region more than 0.6 dB area.

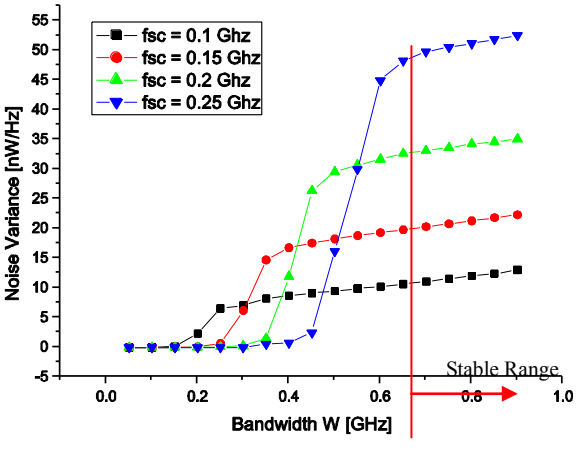


Fig. 3. Calculation results of the noise variance within the narrowband FM bandwidth for the subcarrier frequency at 0.1, 1.5, 0.2, and 0.25 GHz

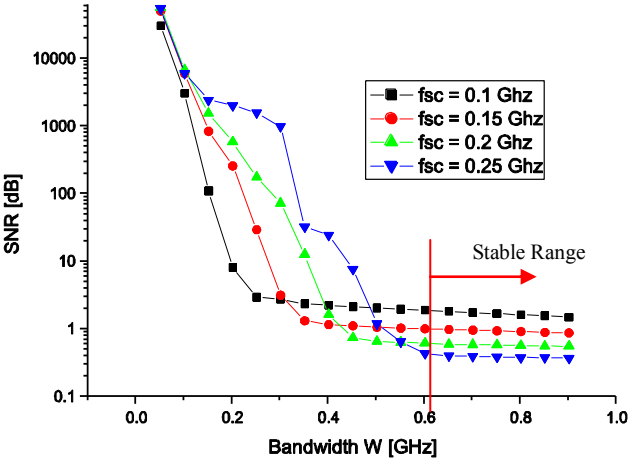


Fig. 4. Calculation results of SNRs for the subcarrier frequency at 0.1, 1.5, 0.2, and 0.25 GHz

In order to plot the calculated performance curve, we substitute the calculated SNR into the formula which expresses the bit error probability for each system. The bit error probability of the *M*-ary PSK scheme is provided by [5].

We have plotted the BER vs SNR curves for the *M*-ary PSK modulators as, BPSK, QPSK, 8-PSK, and 16-PSK schemes in Figs. 5-8 where, we have illustrated the theoretical (calculated) results. The caculated results shows the SNR differences between the modulation schemes are 2.5, 1.1, 4.5 dB at BER = 10^9, respectively. The SNR difference bwtween each sucarrier are 1.9, 2.1. 2.3 dB at BER = 10^9, respectively.

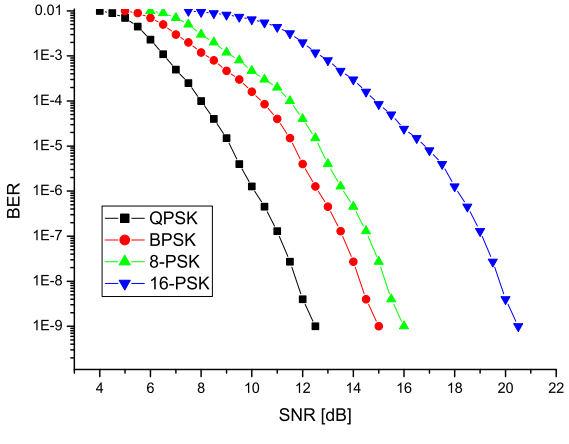


Fig. 5. BER calculation results of *M*-ary PSK Modulations for the subcarrier frequency at 0.1 GHz

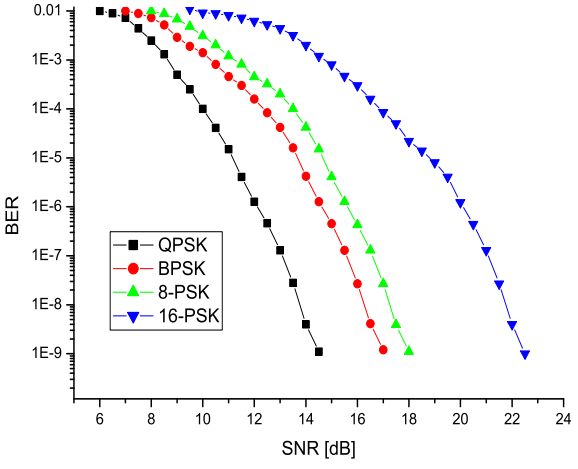


Fig. 6. BER calculation results of *M*-ary PSK Modulations for the subcarrier frequency at 0.15 GHz

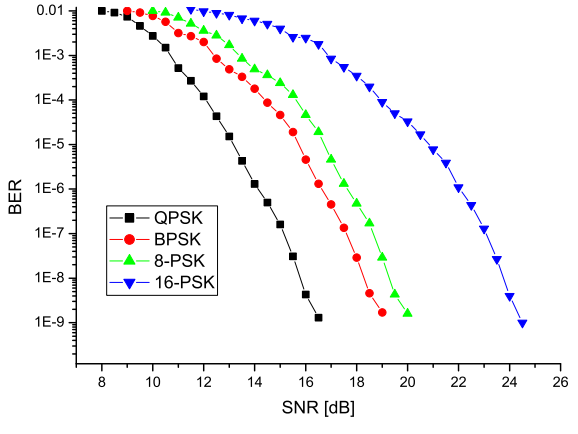


Fig. 7. BER calculation results of *M*-ary PSK Modulations for the subcarrier frequency at 0.2 GHz

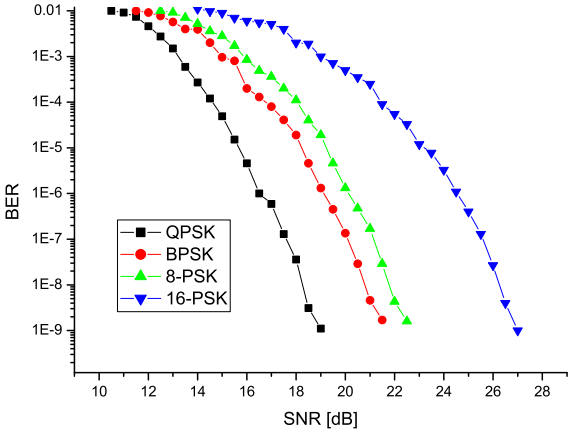


Fig. 8. BER calculation results of *M*-ary PSK Modulations for the subcarrier frequency at 0.25 GHz

4 Conclusions

This paper has described the basic principles and characteristics of the variance at the FM discriminator output for the adjacent sucarrier channels in an optical coherent transmission system using NBFM-SCM. The SNRs for all subcarriers are also stable state at W < 0.6 GHz, which represents even though the subcarrier with a larger noise variance within the FM bandwidth, the performance is compensated by using the digital modulator transmitting lower number of bits. Also, the multiple NBFM-SCM signals can be transmitted by using the WDM. Furthermore, because the shape of FM noise spectrum is colored, the subcarrier modulation with multitone modulation can be used for the even SNRs for all channels.

Acknowledgments. This research was supported by the Research Grants of Catholic University of Daegu in 2011.

References

1. Hui, R., Zhu, B., Huang, R., Allen, C.: Subcarrier Multiplexing for High Speed Optical Transmission. IEEE Jour. Lightwave Tech. 20(3), 417–427 (2002)
2. Teramoto, S., Ohtsuki, T.: Multiple-subcarrier Optical Communication System with Subcarrier Signal Point Sequence. In: IEEE GLOBECOM 2002, Taipei, Taiwan (November 2002)
3. Gross, R., Olshansky, R., Schmidt, M.: Coherent FM-SCM System Using DFB Lasers and a Phase Noise Cancellation Circuit. IEEE Photo. Tech. Let. 2 (January 1990)
4. Olshansky, R., Gross, R.: Subcarrier Multiplexed Coherent Lighwave Systems for Video Transmission. IEEE Jour. Sel. Area Commu. 8 (September 1990)
5. Ziemer, R., Tranter, W.: Principles of Communication. Houghton Mifflin Co. (1990)

A Study on the Effective Lesson Plan of Creative Engineering Design Education for the Creativity Improvement of the Students of Engineering College

An-Na Kang, Sang-Cho Chung*, and Jin-Hee Ku

Innovation Center for Engineering Education & Dep. Mathematics Education,
Mokwon University, Daejon, Korea
anakang25@mokwon.ac.kr, math888@paran.com, jhku@mokwon.ac.kr

Abstract. This study aims to prompt the plan for the creativity improvement through the creative engineering design subjects lesson as the target of engineering certification deepening program, and it was confirmed that 69% students have changed by the survey through the results that 60% students who have negative thinking about creativity before the classes were improved through 'Self-development training program for a positive mindset' and selected design tasks in the beginning of semester. The tasks which have effects on the creativity increase were in the order of brainstorming, egg drop, solar car, and the cause of decrease of creativity were in the order of mismatched team, Negative thinking, Lack of computer literacy.

Keywords: creativity improvement, creative engineering design, engineering students.

1 Introduction

21st century is the knowledge-based society which requires creativity and diversity, and the strength and creativity of individual will be the major national intellectual property. Today, students of engineering college living in the rapidly changing modern society cannot help obtaining ability in order to respond to changes and to cope with new era effectively. In order to actively cope with knowledge-based society, it should shed the undifferentiated education contents and method, and engineering education based on information skills and comprehensive and diverse creativity is necessary.

The creativity education has intensified in the various countries of the world, to achieve this goal, it is materialized by the policy tools that the curriculum is newly set up and the evaluation is strengthened in order to increase quality management of education and social accountability.

In this context, the development and application of curriculum which improves creativity is required, and the creative engineering design education of University is presented as one alternative to realize this purpose.

The creative engineering design education is the subject to operate for improvement of the basic knowledge and problem-solving skills of design traditionally in the

* Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 108–119, 2011.

engineering college, and it may be desirable that the basic design course of the lower grades should cultivate the creative thinking and original design ability so the deepening course of department in the higher grades can be performed well.

In this regard, the case study of developed and operated curriculum of creative engineering design subjects are actively conducting in recent years[1][2], Especially, developed design theme for the creative ability of thinking improvement through the curriculum development for the creative ability of thinking improvement[3]. In addition, it was aimed to research the possibility of applying to lesson in University about the creative troubleshooting teaching-learning method which is designed for self-learning the global social change and required competence in company, in other words, creative thinking power and troubleshooting capabilities for the students through lesson in University, and the possibility of applying to University curriculum about the designed creative troubleshooting teaching-learning method was researched[4][5], and the effects of the troubleshooting progress of the concept design & products design which is conducted through the design education in one semester on the students' self-directed learning ability improvement was examined[6].

The importance of creativity development through the school education was emphasized, and the curriculum design principles and directions were limited by the comprehensive concept.

But the creative engineering design subjects opened on the domestic engineering college is focused on the basic design content which is required by ABEEK[7], so the major purpose is to culture design basic ability rather than to have interest about all conditions for creativity improvement. Therefore, it is hard to find training for creativity improvement through the systematic education program.

On the other hand, suggested that the happy people have higher happy people than unhappy people based on the results that the positive emotion group has higher creativity than the negative emotion group about identifying relationships of happiness and creativity[8]. But it cannot be seen whether the result is caused by creativity promotion of the positive emotion or by the lowering creativity of the negative emotion. Or the person who has creative personality can be regarded as the positive emotion level but it cannot be regarded as low negative emotion level. In addition, as the negative emotion lie sadness is reported that inhibits compromising intellectual judgments estimation, the subsequent studies including the control group should be conducted in order to comprehend the effects of emotion more correctly.

A, N. Kang examined that the creative thinking of students who are thinking that the creativity is the absence is interfered by the lack of confidence and self-deny, so they insisted that overcoming self-confidence is the most important factor of the creativity improvement[9].

Thus, the factor which improves the creativity is defined very diversely and complexly, but the researchers presented that it is very important to inspire the positive mind. Considering this point, it is assumed that the effects of positive emotion is bigger than the negative emotion, but in order to comprehend the effects of emotion more correctly, the effects on the creativity improvement should be identified after aggressive treatment for increasing the positive emotion level.

Therefore, this study aims to answer the following research questions by designing the effective curriculum under the hypothesis that 'the positive mind is help for the creativity improvement.'

(1) Is the education for the students to have the positive mind in the beginning of the semester help for the creativity improvement?

(2) Which design tasks were effective for the creativity improvement? What design tasks was useless?

2 Theoretical Background

2.1 Operating Status of Creative Engineering Design Subjects

Presently, in the domestic engineering college, the departments which conduct the engineering certification program based on KEC2005 engineering certification standards are increasing gradually. The engineering certification program is operated by ABEEK(Accreditation Board for Engineering Education of Korea) aiming the improvements of engineering college education quality, training of excellent engineering researcher, furthermore, the internationalization of engineering education.

According to the subject area certification standards presented by ABEEK, the design related subjects should be organized by all means in order to increase the graduate's adaptation field, especially, the main purpose of the creative engineering design subjects is to develop creativity in the lower grades by the basic design subjects[7]. Therefore the basic design subjects requires to compose the learning content in order to perform the general design subjects of general design subjects by culturing of the basic knowledge and skills of the creative engineering design through the subjects. The specific contents are as follows.

- It should be composed in order to understand the contents about the systematic design course including the meaning of engineering design, design composition factor and practical limit factor etc.
- It should be composed in order to culture the basic ability culture of engineering design, in order words, the creative troubleshooting ability, teamwork ability and communication ability etc.
- It should be composed in order to use in the general design step for the factor design as deepening course by culturing the systematic design course through hands productions, necessary basic ability for design etc.

The operation current state of the basic design subjects such as curriculum contents, teaching materials, design project etc. was examined by selecting 3 schools with experience of producing curriculum contents, teaching materials, design project for the operation of effective creative engineering design subjects.

In the case of C University in South Chungcheong provinces, the basic theory about engineering designs such as the definition of the problem, idea generation, idea evaluation, idea analysis and concept design commonly in the curriculum purpose, teaching materials, class content was learned in almost all subjects which operate engineering certification program, and it is composed to perform the design practice through the design cases. Most of the design project which experience the specific engineering design course are conducting once in the first semester, in addition, the

design project in the idea dimension like 'To improve discomfort in life" which the creative idea is important may be progressed more.

In the case of Y University in North Gyeongsang Province, there are little bit differences of the lecture contents between the faculties, but 'engineering introductory design' which is developed by main teaching materials is used as the common teaching materials, and teaching materials contents teach the theories about engineering design such as definition of the problem, idea creation, idea evaluation, idea analysis, concept design and detail design etc. and perform the design tasks like C University.

In the case of S University in Gyeonggi Province, there were differences of operation method and syllabus between the most of the schools. In the case of S University, the amalgamation design educations of the design social science perspective such as inducing a concrete understanding of the consumer perspective & requirements of major customers, communication between team members, coordination, and improvement of teamwork for the team performance ability improvement etc. were conducted based on the design education research by the Creative Design Institute as the institute affiliated with University. It is operated through the common teaching materials in addition to common syllabus, and it is changed into the institute organization depending on the design project, so it is delivered to each faculty before the start of lectures.

As described above, the creative engineering design education which is operated in the most of the Universities currently has purpose to culture the creative and ingenious design ability for engineering college students but the systematic training for the systematic training which should be followed by the subjects operation of the most of the Universities is lacking. The students will be able to evaluate and design the practical engineering problem by culturing the practical engineering of the realistic context by applying the creativity and problem-solving ability as the creative thinking techniques and procedures.

2.2 Education for Creativity Improvement

The products and services never envisioned have emerged caused by the rapid changing market, technological change and Global competition. In this environment, the 'creativity' became a buzzword, constantly creative techniques and products should be presented in the market for the companies to survive above all, so the excavation of talent with excellent creativity became important issues.

W. Lim[10] analyzed the reason that all people who has knowledge of professional level cannot show the creative performance despite its important is that their knowledge promotes fixation phenomenon. The fixation means the phenomenon to interfere the cognitive operation in the situation that the memory, problem-solving course and creative thinking are expressed by dictionary knowledge, and these constraints have strong effects on the generation of creative idea of individual and the group as well.

Therefore, the creative problem-solving premises expert knowledge of the expert level, and this knowledge promotes necessarily fixation, if it is true that the creative problem-solving is interfered, how the interfere effects involving solving the problem creatively minimized should be examined.

But the reason that the excavation of creativity has complexity is not the fact that simply the various factors are involved, but that is because of the fact the excavation course of creativity is not decided as the simple sum of each factor. As pointed out by Lubart and Guignard[11], the recent studies on the necessary factors for excavation of creativity has interest in combination of cognitive factors, synchronous factors and environmental factors.

In addition, M. H. Shin[8] explained that two emotions of joy and sorrow have effects on the positive emotion and negative emotion in the study of identifying the relations between happiness and creativity. In addition, the emphasized point by divergent thinking and side thinking commonly is expansion of thinking, and this expansion of thinking is regarded as being induced by the positive emotion. According to this opinion, it is described that the negative emotions such as fear, anger, disgust causes the psychological result to shrink repertory of psychological result temporarily by acting in the specific method like to run away, to attack, to expel.

This study aims to present systematic lesson operation plan which can help the systematic lesson operation plan based on training to understand and removal factors inhibiting creativity for learning strategies related creative design performance for culture of creative design ability.

3 Main Subject

3.1 Materials and Methods

This study is the examination result of two times of before and after the classes based on 122 students with Department of Construction Engineering (64 students), Department of fire Service Administration(5 students), Department of Electrical Engineering(53 students) who attended the lecture of creative engineering design subjects opened as engineering certification subjects of the first semester in 2011 in A University in Daejeon area.

The questionnaire was used by self-production, and the survey contents were composed by three background variables related (optional) questions about departments, home economic power and happiness indices etc. and four questions about the presence of creativity, whether increase or not, creativity increase tasks and decreased factors, etc. The issues were examined by the survey examination result and it were used as the part of lesson of the creative engineering design introductory course의 lesson. It was plan to induce the self potential benefits of about 60% of students with negative thinking by the results about cause of the presence of creativity before the lesson, and the participation and creativity of design project lesson after midterm were improved by conducting lesson which can change the negative mind into positively after introducing self development project for early three weeks of 16 weeks lessons.

The correlation between each data was analyzed for examining relations about positiveness degree and creativity improvement by using survey data through the questionnaire. In addition, T- test and evaluation comparing the response materials conducted statistical processing by using Microsoft Excel and PASW Statics 18.

3.2 Weekly Lecture Contents

The weekly lecture contents expected that the aggressive positive thinking and open thinking help the self creativity improvement of students more than the negative thinking, so it was focused on maximizing the creativity improvement of students through the curriculum such as self-discovery, teamwork, creativity recognition factor training program, brainstorming, problem-solving ability improvement, completion of design tasks etc. for 16 weeks.

In beginning of semester, it was applied by re-adjustment into 5 steps in order to operate students' creative engineering design education efficiently based on Dr. Roef Smith의 "The 7 Levels of Change:" by Dr. Roef Smith, the convergent thinking by Guilford[12], and feature of divergent thinking as self-development training learning for 3 weeks[9][13][14]. There were 5 steps of self-transformation such as discovery→ development → cutting → differentiation → challenge(confidence)

1step: discovery → self-recognition factor discovery, 2step : development → beginning of change, 3step : cutting → negative factor removal, 4step : differentiation → differentiation between myself and others, 5step : challenge → expansion of creative thinking ability.

The design tasks project of other 13 weeks are as follows.

- Team composition: the design project team is composed voluntary by 5~6 members considering the personal personality and grads.
- Selection of design tasks: the design project tasks were selected by reflecting the preferences of students referring the survey result.
- Progressing of project: the given tasks were completed by presenting definition and solving plan of problem through brainstorming by each team.
- Project evaluation: the evaluation was composed by dividing into tasks by individual and tasks by team after making the standards plan of the design composition factor and realistic constraints.

Syllabus is as <Table 1>.

Table 1. Syllabus of creative engineering design subjects

Week	Lecture contents
1	Subjects introduction & survey research conducted
2	Self-development training program for a positive mindset 1
3	Self-development training program for a positive mindset 2
4	Self-development training program for a positive mindset 3
5	Approaching method of problem-solving, theoretical background and using synesthesia
6	Team projects execution – Shaking off worries far away, comments deriving comments opinion through brainstorming

Table 1. (*continued*)

7	Team projects execution – creation of the value of elastic cord, scamper(SCAMPER)techniques
8	Team projects execution – free-fall of eggs
9	Team projects execution – presentation of tasks
10	Team projects execution – Creating speaker
11	Team projects execution – Create a paper bridge 1
12	Team projects execution – Create a paper bridge 2
13	Team projects execution – Create a solar car 1, Scientific deliverables to improve the discomfort in life
14	Team projects execution – Create a solar car 2, Scientific deliverables to improve the discomfort in life
15	Team projects execution – Create a solar car 3, Scientific deliverables to improve the discomfort in life
16	Team projects execution – Contest

3.3 Result of Study

As results of survey basic examination, the correlation with creativity about home level and happiness degree of home, self positiveness degree showed 0.56 as examination of internal consistency(α value of Cronbach), so it can be seen that there is reliability of degree survey. The creativity and correlation between home level and happiness degree of home were shown as –0.36~0.14 degree, so it can be seen there is no correlation, but the person correlation with creativity degree, self positiveness degree and creative engineering design subjects before the classes were 0.25(level of significance<0.05) and it showed 0.41(level of significance<0.01) as the creativity degree with creative engineering design subjects after the classes, so the correlation was higher than before the classes.

The survey result of Likert 2 points scale evaluation type with ‘I have creativity’ or ‘None’ before and after the 16 weeks classes are as <Table 2> and <Table 3>.

<Table 2> showed that the survey(‘Yes’ is 1 point, ‘None’ is 2 points) about ‘Do I have creativity’ as the examination results of 76 respondents about survey before and after the classes by the paired sample T test(level of significance 0.01). The average of the reply subjects before the classes is 1.59, there are many ‘None’ in thinking (average point as 1.5), but the average respondents after the classes is 1.29, so ‘No creativity’ means the 0.30 degree decreased opinion. It shows that the negative mindset was changed into positively by the effects of ‘Self-development training program for a positive mindset’ for 2-4 weeks in the beginning of semester.

Table 2. T test result of paired sample before the classes

76 persons	Average	Standard deviation
Before the classes	1.59	0.50
After the classes	1.29	0.46
Differences before and after the classes	0.30	0.52

<Table 3> showed the examination result of Likert 2 points scale evaluation about 86 survey respondents before the classes and 122 survey respondent after the classes. It can be seen that the respondent ratio with 'I have creativity' was 39.5% before the classes, but it increased with 68.9% after the classes. Especially, Department of Electrical Engineering increased from 25.9% before the classes into 71.7% after the classes. It is judged that had confidence by assessing to problem-solving plan through the proper selection and brainstorming of design tasks of design tasks by increase cause.

Table 3. Survey examination of creative engineering design introductory before and after the classes – respondents who answered that 'I have creativity'

Departments Before & after	Building		Fire fighting		Electron		Whole	
	Personnel	%	Personnel	%	Personnel	%	Personnel	%
Before the classes (86 persons)	26	48.1	1	20	7	25.9	34	39.5
After the classes (122 persons)	43	67.2	3	60	38	71.7	84	68.9

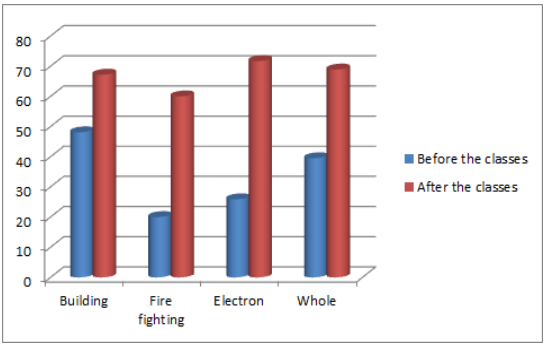


Fig. 1. The graph of survey examination of creative engineering design introductory before and after the classes – respondents who answered that 'I have creativity'

<Table 4> showed the examination result by Likert 5 points scale(Greatly increased, Increased, Normal, Decreased, Greatly decreased) about increase of self creativity through the students after the classes, the 'Greatly increased' was 8.4%, 'Increased' was 58.8% as responses, so 67.2% answered that the self creativity is increased. The students who answered that decrease were only 2 (1.7%) among whole students, so it can be seen that there are effects on the lessons about creativity improvement.

Table 4. Whether creativity increase or not after creative engineering design introductory lesson

Departments Increase	Building		Fire fighting		Electron		Whole	
	Personnel	%	Personnel	%	Personnel	%	Personnel	%
Greatly increased	6	9.7	0	0.0	4	7.7	10	8.4
Increased	32	51.6	2	40.0	36	69.2	70	58.8
Normal	23	37.1	2	40.0	12	23.1	37	31.1
Decreased	1	1.6	1	20.0	0	0.0	2	1.7
Greatly decreased	0	0.0	0	0.0	0	0.0	0	0
Total	62	100	5	100	52	100	119	100

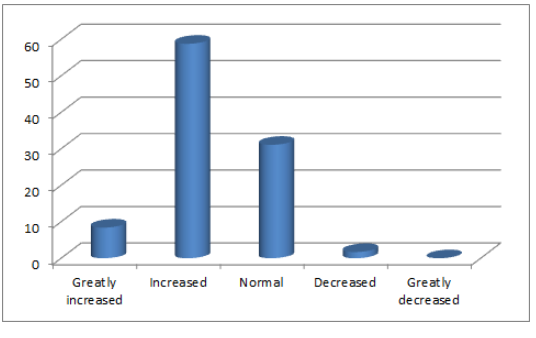


Fig. 2. The graph of Whether creativity increase or not after creative engineering design introductory lesson

In order to verify the effects of creativity before and after the classes, one of the methods to measure the creativity of survey should perform the tasks type as "Draw the subjects by using circle as many as you think" and the result before the classes was 2.81(standard deviation 2.81) but it average 0.98 increased with 3.79(standard deviation 3.16) after the classes, so it can be said that there were effects on improving creativity class.

The survey results of selection for 5 types per one person about the tasks helping the creativity among the contents conducted in 16 weeks lessons is <Table 5>. The

brainstorming was topped with 13.8% as response in the several tasks, and followed by egg drop and solar car in the order of. There were few differences by departments, and the Department of Electrical Engineering considered that the creativity will be improved by the self-development training program as 2nd priority.

Table 5. Class content of cause of self creativity increase

Departments Class content	Building		Fire fighting		Electron		Whole	
	Personnel	%	Personnel	%	Personnel	%	Personnel	%
1. Brainstorming	34	17.7	3	16.7	20	9.9	57	13.8
2. Egg drop	28	14.6	1	5.6	25	12.4	54	13.1
3. Solar car	20	10.4	2	11.1	29	14.4	51	12.4
4. Value creation	23	12.0	2	11.1	20	9.9	45	10.9
5. Harmony of team members	21	10.9	3	16.7	19	9.4	43	10.4
6. Self development program	13	6.8	1	5.6	25	12.4	39	9.5
7. Life sciences artifacts	9	4.7	1	5.6	23	11.4	33	8.0
8. Shaking off worries	14	7.3	0	0	16	7.9	30	7.3
9. Using synesthesia	8	4.2	1	5.6	16	7.9	25	6.1
10. Scamper techniques	13	6.8	3	16.7	8	4	24	5.8
11. Etc.	9	4.7	1	5.6	1	0.5	11	2.7
Total	192	100	18	100	202	100	412	100

<Table 6> is the results that five things selected per one person about the creativity decrease factors about students. The counts (57 times) selected as decreased are very smaller than the counts(412 times) selected as increased, but it was thought that the inconsistency of the team members and negative thinking decreased creativity most, and followed by the cause as the lack of computer literacy. As the complaints elements of team, it was judged that have effects on teamwork depending on attendance & personal interest degree about participation. Therefore, the study for preparing the specific plan for resolving complaints must be followed.

In the case of the architecture departments, there are lack of self development and complains about lesson method, so the parts about the lesson method needs to be improved.

Table 6. Occurrence factors of cause of self creativity decreased

Departments Occurrence factors	Building		Fire fighting		Electron		Whole	
	Personnel	%	Personnel	%	Personnel	%	Personnel	%
1. Mismatched team	2	16.7	2	20	5	14.3	9	15.8
2. Negative thinking	2	16.7	1	10	6	17.1	9	15.8
3. Lack of computer literacy	1	8.3	1	10	6	17.1	8	14.0
4. Lack of life experience	1	8.3	2	20	4	11.4	7	12.3
5. Lack of self development	2	16.7	1	10	3	8.6	6	10.5
6. Lack of lesson	1	8.3	1	10	3	8.6	5	8.8
7. Disappointment about lesson method	2	16.7	0	0	3	8.6	5	8.8
8. Lack of self ability	0	0	1	10	4	11.4	5	8.8
9. Etc.	1	8.3	1	10	0	0	2	3.5
10. Difficulty adapting school	0	0	0	0	1	2.9	1	1.8
Total	12	100	10	100	35	100	57	100

4 Conclusions and Recommendations

In this study, it was confirmed that the students' creativity which had self negative mindset about creativity before the classes of the creative engineering design introductory course curriculum increased through the survey by designing the curriculum inducing the self potential benefits through the 'Self-development training program for a positive mindset' for 3 weeks in the beginning of semester. The general results are as follows.

First, 40% students before the classes were negative about creativity, but 69% students answered that they have creativity after the classes.

Second, as one of the causes that the creativity increased is the 'Self-development training program for a positive mindset' for 3 weeks in the beginning of semester.

Third, the tasks which had effects on the creativity increase among design projects were shown in the order of brainstorming, egg drop, solar car. It is considered that the definition and solving method of problem was expressed confidently through brainstorming as the major cause.

Fourth, as the cause to decrease creativity, it was shown as the mismatched team, negative thinking, lack of computer literacy, so the methods to compose the smooth teamwork should be promoted, and the various teaching method which increase confidence is needed to be developed.

Finally, this study confirmed that the creativity improvement through the given confidence is available in somewhat degree, but specifically, there is necessity to research which degree can be improved through the test by the group. In addition, there are differences of the design tasks contents with high affinity depending on department characteristic, so it is considered that there is necessity to adjust the creative design subjects by departments.

References

[1] Lee, E.H., Bae, W.B.: Analysis of Design Pattern of Engineering Design Curriculum. Journal of Engineering Education Research 10(3) (2007)
[2] Lee, J.S., Min, B.K., Yoon, W.S., Hahn, J.W., Jung, H.I.: A Base Course of Creative Mechanical Engineering Design Emphasizing Experience Based Learning. Journal of Engineering Education Research 11(2) (2008)
[3] Kim, L.H., Lee, B.S.: Development of Introductory Engineering Design Course to Improve Creative Thinking Ability. Journal of Engineering Education Research 8(3) (2005)
[4] Park, C.S., Park, S.H., Jeong, S.Y.: Research on Applicability of Teaching-Learning Methods for Creative Problem Solving to a Course in University. Journal of Engineering Education Research 13(1) (2010)
[5] Han, J.Y.: Improvement of Self-Directed Learning Ability through Engineering Design Education. Journal of Engineering Education Research 14(1) (2011)
[6] Kang, H.S., Kim, C.H., Lee, J.S.: The Principles and Orientations of Curriculum Design for Developing Creativity. The Secondary Education Research 51(2), 1–39 (2003)
[7] ABEEK(Accreditation Board for Engineering Education of Korea), http://www.abeek.or.kr
[8] Shin, M.H., Koo, J.S.: The Relationship between Happiness and Creativity. Korean Journal of Social and Personality Psychology 24(3) (2010)
[9] Kang, A.N., Nam, S.Z., Ku, J.H.: Effectiveness of the Class: Principles of Creative Engineering Design based on Self Development Leadership(SDL). In: ICCC (2010)
[10] Lim, W.: The improvement of creativity through curriculum education. In: Creativity Improvement Workshop (2009), http://www.omegaedu.co.kr
[11] Lubart, T., Guignard, J.: The generality-specificity of creativity: A multivariate approach. In: Sternberg, R.J., Grigorenko, E.L., Singer, J.L. (eds.) Creativity: from Potential to Realization, pp. 43–56. American Psychological Association, Washington, DC (2000)
[12] Von Oech, R.: A Whack on the side of the head. Warner Communications, New York (1983)
[13] Moon, Y.R.: Multiple Intelligences: New Horizons. Woongjin Jisik House (2000)
[14] Guilford, H.P., Hoepfner: The Analysis of Intelligence. McGraw-Hill, New York (1971)

Frameworks for Multi-purpose U-Health Care Interface

Haeng-Kon Kim

Dept. of Computer Information & Communication Enginerring,
Catholic University of Deagu, Korea
hangkon@cu.ac.kr

Abstract. Rather than modeling the user-interface as a set of static structures and mappings, the u-health care UI should be modeled as a set of design preferences. Preferences are frequently many-to-one or many-to-many relationships that elude conventional u-health care UI modeling, which has largely focused on one-to-one mappings. There are network compatibility problems among the u-health devices as their different protocols and software which are manufactured depending on the characteristics of companies.

In this paper, a spectrum of preference relations is described, and a new syntax for modeling preference is proposed to design the u-healthcare UI. This spectrum extends from simple one-to-one bindings to complex design guidelines that can be structured together to implement decision trees. This new representation allows decision trees to be tightly integrated into the user-interface model itself, enhancing their flexibility and power. we also analyzed ISO/IEEE 11073 which is standard of information exchange among device agent and service manager and designed EMDI(Electronic Medical Data Interchange)interface for the improvement of compatibility in u-healthcare service field. The proposed interface structure can manage and exchange all of service data among device agent and application system on the integrated gateway environment.

Keywords: UI, u-healthcare, interface, OSGi, Service Model.

1 Introduction

While a general-purpose framework for representing such mappings might be desirable for the flexibility that it offers, such a framework would add considerable complexity to the representation language and impose a significant burden on user-interface modeling tools. One specific phenomenon that commonly eludes description by one-to-one mappings is preference. Preference relations allow for a more flexible and adaptable specification of the u-healthcare user-interface rather than specifying exactly how the u-healthcare UI must appear, the designer can specify what would be preferred, and in which situation. This freedom is particularly important for user-interfaces that must run on heterogeneous devices [1,2] since the contexts of use vary and may even change at run-time. Preference modeling is also critical if interaction is

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 120–130, 2011.

to be dynamically customized for the user; preference relations can be used to show how the user-interface should change in relation to the user model. On the other hands, u-Healthcare technology can be used for diagnosis and medical treatment at any time, any place connected with networks for the aged and patients. It is expected that the u-healthcare can solve the medical cost problem thru managing the chronic disease. Rapid development of information and communication technology also supports their health concern via remote diagnosis, emergency medical treatment, self monitoring and so on. Thus the communication system and interface system is considered so important element. Figure 1 shows individual healthcare technology.

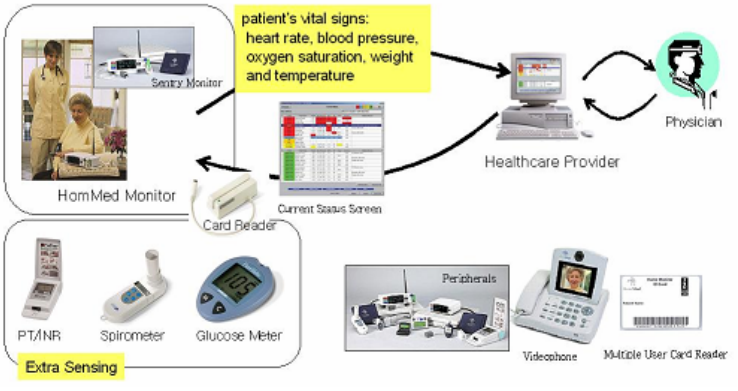


Fig. 1. Elements of healthcare technology

However, each individual healthcare device contains unique protocol and related software of each different company and those protocols are not opened to the publics. Thus, there are a few data compatibility among such devices and the standardization to solve the problem become very important issue. Nevertheless, the standardization has hardly done as increasing cost of the measuring- instruments and software. For the more, medical data of non standard format can't be inter-operated or integrated with EHR(Electronic Health Record) without global standardization. It is a record in digital format that is capable of being shared across different health care settings, by being embedded in network-connected enterprise-wide information systems.

In this paper, a spectrum of preference relations is described, and a new syntax for modeling preference is proposed to design the u-healthcare UI. This spectrum extends from simple one-to-one bindings to complex design guidelines that can be structured together to implement decision trees. This new representation allows decision trees to be tightly integrated into the user-interface model itself, enhancing their flexibility and power. we also analyzed ISO/IEEE 11073 which is standard of information exchange among device agent and service manager and designed EMDI(Electronic Medical Data Interchange)interface for the improvement of compatibility in u-healthcare service field. The proposed interface structure can manage and exchange all of service data among device agent and application system on the integrated gateway environment.

2 Related Works

2.1 Abstract Preferences

Bindings, simple preferences, and ordered preferences are all concrete preferences, because the targets are specified directly. Rather than specifying the element that is preferred, a designer may wish to instead specify the characteristics of the element that is preferred. For example, a designer may prefer whatever presentation element requires the least number of clicks, or whatever dialog structure imposes the least cognitive load. This kind of preference is referred to as an abstract preference. Abstract preferences have an additional feature—a set of criteria. The criteria determine which target is selected; in the example above, the number of clicks required is the criterion. For an abstract preference, the targets are possible selections, and criteria determine which target is actually chosen. Needless to say, an abstract preference is required to have more than one target. There are two types of criteria: preferential and logical. Preferential criteria specify those characteristics that cause one target to be preferred. For example, a preferential criterion might say to choose the dialog structure with the least complexity. Conversely, logical criteria specify what kinds of targets are allowed. If there is more than one preferential criteria, then each criterion must have a priority to indicate how important it is relative to the others. Logical criteria need no priority; if the value of the feature under consideration by a logical criterion is in violation of the criterion, then the associated target is ruled out. When used in combination, preferential and logical criteria can allow for very complex preference specifications. For example, consider the following specification: "For task model Tm2 and domain model Dm1, user U4 likes presentation elements that require few clicks, take up little screen real estate, and have lots of colors. The criteria of having lots of colors is most important, followed by the number of clicks. But if amount of space occupied is too small, it won't be visible, so disallow it altogether." In this situation, Tm2, Dm1, and U4 are the conditions. The set of presentation elements under consideration are the targets. There are three preferential criteria: number of colors (more is preferred), number of clicks (less is preferred), amount of space occupied (less is preferred). In addition, there is one logical criteria: the amount of space occupied must be greater than some constant.

2.2 Multi-agent

A multi-agent system (MAS) is a system composed of several agents, capable of mutual interaction. The agents are considered to be autonomous entities such as software programs or robots. Their interactions can be either cooperative or selfish. That is, the agents can share a common goal or they can pursue their own interests. In [2] propose intelligent information retrieval (IIR) agents as a solution to the information overloading. It identifies the desirable features of an IIR agent, including intelligent search, navigation guide, auto-notification, personal information management, personal preferred interface, and tools for easy page-reading. In [3] discusses two distributed information filtering approaches that are distributed with respect to knowledge or functionality, to overcome the limitations of single-agent

centralized information filtering. Large-scale experimental studies involving the well-known TREC data set are also presented to illustrate the advantages of distributed filtering as well as to compare the different distributed approaches.

2.3 Global Standardization Trends of IEEE 11073 PHD

IEEE 11073 PHD is the standardization for the exchanging information among private medical remote devices and manager. Those managers are mobile phone, computer, set top box, information collector and so on. Figure 2 shows general structure of private medical system which is using in u-health.

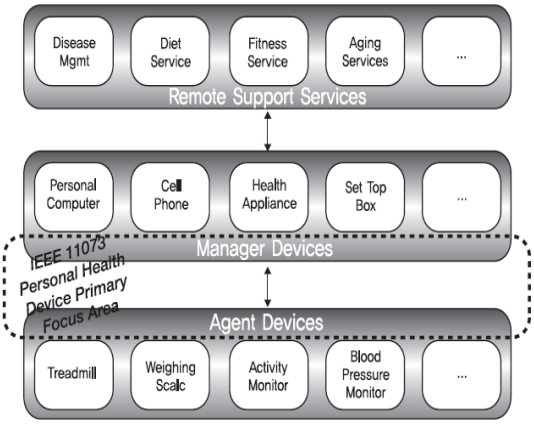


Fig. 2. Architecture of private healthcare devices

Private healthcare devices can collect patient's information and transmits to the manager such as mobile phone, computer, medical facilities. Manager can transmit the data to the remote supporting center for the detailed analysis. Normally, the communication channel among private healthcare devices and manager is carried out by point-to-point connection. Manager can communicate with multiple private devices using point-to-point protocol. 11073 PHD standard model defines protocol structure, private healthcare device and CE and is constructed as in figure 3 models as follows.

Domain Information Model(DIM): It is Object-Oriented Model and defines object of private healthcare device. Each object has attributes more than one and each attribute shows measurement data or function of device.

Service Model: It defines access method between private healthcare device and manager. It access as defined in DIM whenever accessing the data. Service commands include GET,SET, ACTION, Event Report etc.

Communication Model: It defines network communication protocol by point-to-point connection. Generally, MDER(Medical Device Encoding Rule)is used as the coding of DIM in IEEE 11073 PHD. Figure 3 shows 11073 PHD standard model structure as described above.

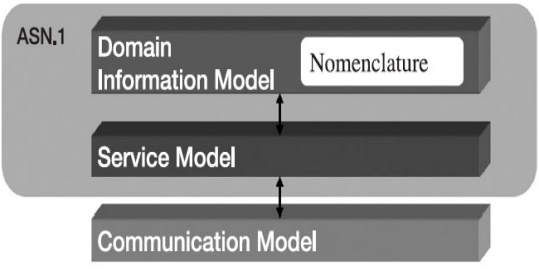


Fig. 3. IEEE 11073 PHD standard model structure

IEEE 11073 PHD includes protocol stack for the connection of private healthcare device and manager. IEEE 11073 PHD protocol stack is structured as figure 4 and divided into 3 level as follows.

Transport Layer: IEEE 11073 PHD allows various transmission methods without detailed physical definition.

Optimized Exchange Protocol: This level is most important part in IEEE 11073 PHD. It provides supporting foundation of various private healthcare device. Application service provides management of connection between private healthcare device and manager in addition to the reliable protocol. Also it convert to Medical Device Encoding Rule(MDER), standard Binary ER(BER), Packet ER(PER) from ASN. ata structure.

Device Specialization: On the top of IEEE 11073 protocol stack, there is device specialization layer which describes special detail item related with 8 private healthcare devices. This describes object and attributes of each private healthcare device. Figure 4 shows IEEE 11073 PHD protocol stack.

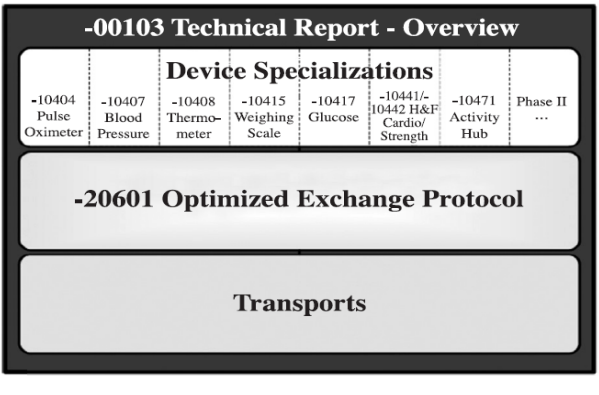


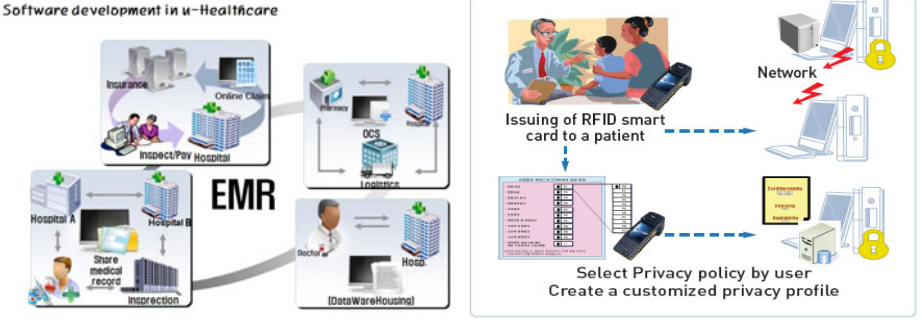
Fig. 4. IEEE 11073 PHD protocol stack

3 Design of U-HealthCare Interface Using User Preference

3.1 Design Guidelines

It has been assumed that the criteria apply only to the u-health care interface targets. That is, distinctions can be made between the u-health care targets only on the basis of characteristics of the targets themselves. It is possible to say, for example, "U4 prefers the least complex task tree," but it is not possible to say, "If U4's experience level is less than intermediate, then use presentation element P3." To do this, we need to separate the list of u-health care targets from the criteria. Preference relations where the criteria do not necessarily apply to the targets are called design guidelines. Design guidelines are the most abstract and complex preference relations that we will attempt to model. For this class of preference relations, an additional feature must be considered: a list of objects. The criteria refer to features of each object. For design guidelines, all criteria must be logical. If the value of every criteria is true for every object, then a mapping is created between the conditions and each of the targets.

The u-healthcare target of a design rule could itself be another preference relation. For example, suppose we wanted to model the following preference: "if U4's experience level is less than intermediate, then use the task tree with the least amount of complexity." Once again, U4 is the object, and experience level is the criterion. But the target is a preference relation – specifically, an abstract preference. That abstract preference says to choose the task tree with the least amount of complexity. So for the abstract preference, the target is a list of task trees and the criterion is the amount of complexity. The condition of the abstract preference is simply the design guideline that targeted it. The interface environment is characterized as Figure 5(a) and (b). It shows Device interface between agent and manager, also it shows the interface between manager and EDI. IEEE P11073-20601 which compose data exchange standard on application layer define logical session, data transmission, session disconnection on the view point of object.



(a) u-halth care Interface domain (b) Usecase Stake holdes

Fig. 5. U-healthcare Interface Environment

3.2 Device Specialization Standard in Our Works

We apply the IEEE 11073 PHD standard depending on Device specialization in our works as followings;

- 11073-10404: Device specialization-pulse oximeter
- 11073-10407: Device specialization-Blood Pressure
- 11073-10408: Device specialization-Thermometer
- 11073-10415: Device specialization-Weighing scale
- 11073-10417: Device specialization-glucose meter
- 11073-10441: Device specialization-Cardiovascular fitness and activity
- 11073-10442: Device specialization-Strength fitness equipment
- 11073-10471: Device specialization-Independent living activity hub

We organizes these standard device in our works are shown as in Figure 6.

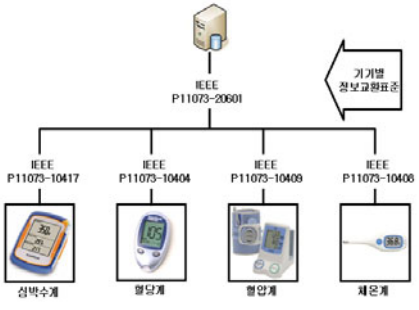


Fig. 6. Device specialization standard

3.3 Device Interfaces

Figure 7 shows device interface architecture. Device interface can make connection application and medical devices. It can also map application and user in addition to the manipulation of device. Device API provides connection of application and Device Driver Manager. Device Driver supports the connection of Device Driver Manager and device.

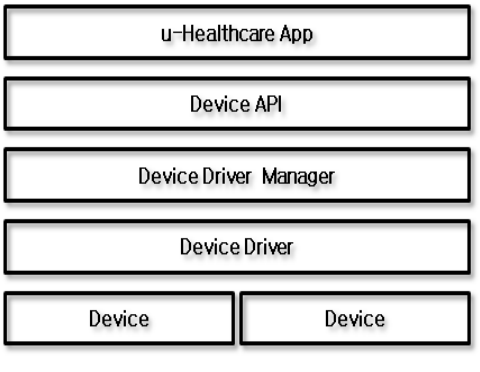


Fig. 7. Device Interface Architecture

4 Design of Function Module Based on OSGi

4.1 EDI Converter Based on OSGi

We designed EDI Protype that is contained in EDI Converter and is XML form for the efficient connection of various device and application system. This EDI Protype can perform the function of data exchange among various device and application system depending upon each device type. Figure 8 shows EDI Converter Architecture

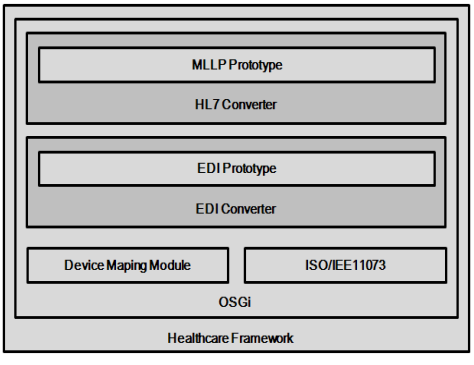


Fig. 8. EDI Converter Architecture

4.2 Service Model

Service Model defines executive concept for the data exchange among agent and manager. For example, object access service of blood pressure gauge is used for accessing the object defined in domain information model of blood pressure gauge. It supports the following services.

- GET service: It is used whenever manager get the object attributes of MDS(Medical Device System)
- SET service: : It is used whenever manager set the object attributes of MDS.
- Event report service: It is used whenever agent sends the measurement data and configuration report to the manager.
- Action service: It is used whenever manager call the supporting action of agent.

Figure 9 shows service model of blood pressure gauge as a service model.

4.3 Modeling Multi-purpose U-HealthCare Interface

In Fig. 10, we classify the candidate class with domain actors map to agents and their functions/goals map to tasks. The concrete domain actors (and the mediator) are implemented directly as agents. In the case of designing UI, for example, this means that they would extend the our frameworks for their own application area in the future.

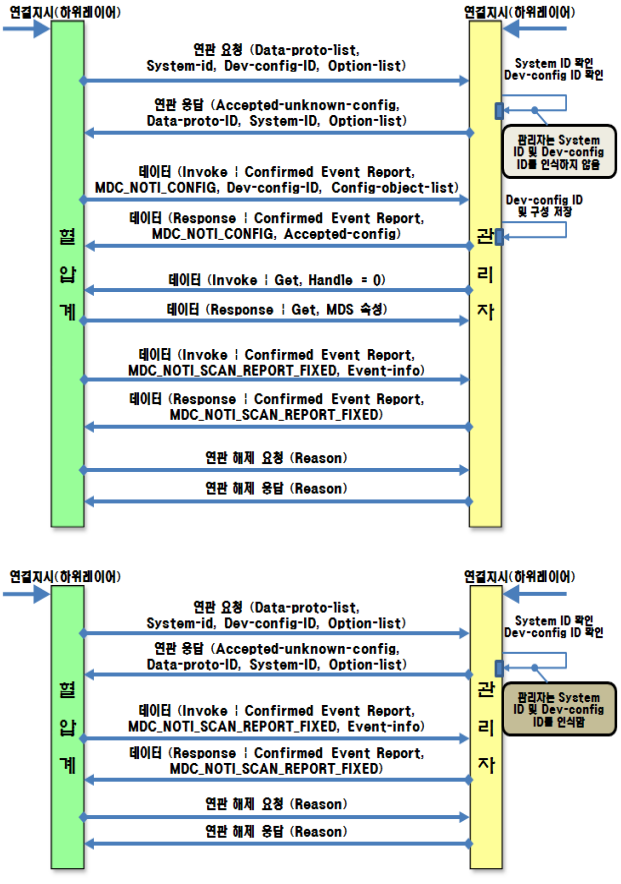


Fig. 9. Service model of blood pressure gauge

Table 1. Mapping of Architectural Elements to our Frameworks

Generic Architecture Element	Mapping
Abstract Domain Actor	Interface, Listener Task
Concrete Domain Actor	Agent
Goal	Task
Domain Entity	Class Module, XML Document

In order to provide the appropriate abstract behavior, the goals and functions/responsibilities table 1 of the abstract domain actors are implemented as Java interfaces(e.g. DiagnosisProducerAbility and DiagnosisConsumerAbility in Fig. 10), but also as descendents of the Task class (Fig. 11). This duality is only necessary to coordinate the entities collaborating to achieve the goal. The test data collected for each patient is stored in a database as in figure 12, overall running environment for

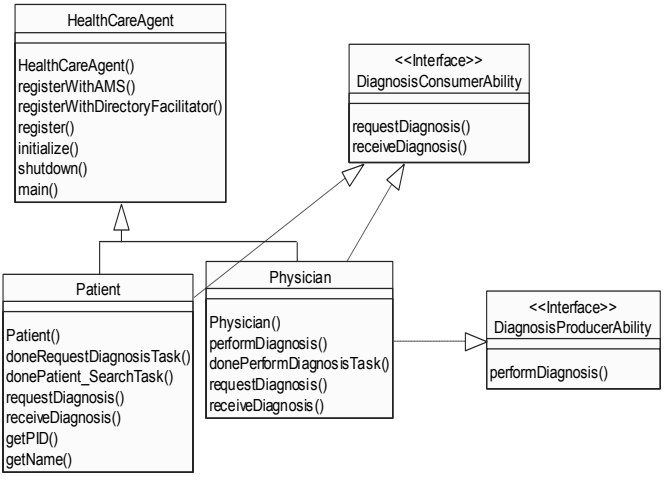


Fig. 10. Class diagram for multi-purpose u-health interface

u-health multi-purpose interface. Each new set of test parameters collected from the various ophthalmic devices) is run through the frameworks, (eventually together with the medical expert opinion and subjective evaluations) and as result the 'assessment of the patent status' is obtained.

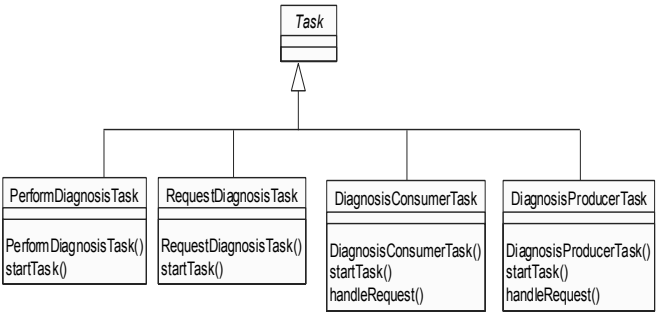


Figure 11. Task diagram for multi-purpose u-health interface

5 Conclusions

In the future, u-health will be further integrated into our daily lives such that we will not perceive its presence consciously, and it will allow us to monitor our health status naturally and continuously. However, even the most highly developed technology can be neglected by consumers if it is not user-friendly. Therefore, u-health needs to fully reflect the required services to obtain maximal participation and behavioral change. U-health market will be reorganized as healthcare device and platform in the near future. Non-standard private healthcare device does not support compatibility among other devices. Thus it disturbs u-health market being activated. We designed EDI

interface structure for the interoperation of device and service application under the OSGi environment which support home healthcare connection. It includes EDI prototype in EDI converter on OSGi and service model which support 4 services. We will implement integrated gateway using this interface structure.

Acknowledgement. This work was supported by the Korea National Research Foundation (NRF) granted funded by the Korea Government (Scientist of Regional University No. 2011-0013259).

References

1. Eichelberg, M., Aden, T., Riesmeier, J., Dogac, A., Laleci, G.: A Survey and Analysis of Electronic Healthcare Record Standards. ACM Computing Surveys 37(4) (December 2005)
2. A Roadmap for Interoperability of eHealth Systems in Support of COM 356 with Special Emphasis on Semantic Interoperability, http://www.srdc.metu.edu.tr/webpage/projects/ride
3. Hapi HL7 SDK, http://hl7api.sourceforge.net/
4. Unlocking the Power of Health Information, http://www.hl7.org/
5. Oh, H.-S.: u-Healthcare Technology and trand of Standard. IT Standard & Test Journal (112)
6. Korea Food & Drug Administration: Standard of U-Health Medical Devices Guideline (2007)
7. Kim, C.-S.: The Trends and Prospects of Health Information Standards: Standardization Analysis and Suggestions. Department of Radiological Science, College of Health Sciences, Catholic University of Pusan (2008)
8. Orguna, B., Vub, J.: HL7 ontology and mobile agents for interoperability in heterogeneous medical information systems. Computers in Biology and Medicine 36, 817–836 (2006)
9. Kim, T.S.: A Metadata System for HL7 aECG Document Management. Department of Computer & Information Engineering Graduate School, Chongju University (2005)
10. Park, H.: Implementation of Hospital Information System based on HL7 using RFID. Department of Electronics & Communications Engineering Graduate School of Kwangwoon University (2006)
11. Lee, T., Choi, M., Yun, C.: Development And Implementation of System for Delivery of Emergency Patient's Basic Information Between Related Hospitals. Journal of Health Science & Medical Technology 29(2), 67–80 (2003)

Toward New Vision of XLINK

Seifedine Kadry[1] and Ali Kalakech[2]

[1] Faculty of Sciences, Lebanese University, Lebanon
[2] Faculty of business, Lebanese University, Lebanon
skadry@gmail.com, alikalakech@hotmail.com

Abstract. In this article, we present the limitations of HTML hyperlink and how could be solve it by using a new XML based language XLINK. Till now there is neither clear specification nor implementation of this language, a new comprehensive design using UML is proposed.

Keywords: HTML hyperlink, XML, XLINK and UML.

1 Introduction

The XML Linking Language, or Xlink [3, 6, 8], is a markup language used for creating hyperlinks in XML documents. Xlink is a W3C specification that outlines methods of describing links between resources in XML [1, 4, 5, 7, 9, 10, 11, 13, and 14] documents, whether internal or external to the original document. Xlink is an explicit relationship between resources or portions of resources. It is made explicit by an Xlink linking element, which is an Xlink conforming XML element that asserts the existence of a link. There are six Xlink elements, only two of them are considered linking elements. The others provide various pieces of information that describe the characteristics of a link.

Xlink provides a framework for creating both basic unidirectional links and more complex linking structures. It allows XML documents to:

- Associate metadata with a link.
- Express links that reside in a location separate from the linked resources.

An important application of Xlink is in hypermedia systems that have hyperlinks. A simple case of hyperlink is an HTML [2, 12] "a" element, which has these characteristics:

- The hyperlink uses URIs as its locator technology.
- The hyperlink is expressed at one of its two ends.
- The hyperlink identifies the other end.
- Users can initiate traversal only from the end where the hyperlink is expressed to the other end.
- The hyperlink's effect on windows, frames, go-back lists, style sheets in use, and so on is determined by user agents, not by the hyperlink itself.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 131–146, 2011.

These set of characteristics are powerful, but the model that underlines them limits the range of possible hyperlink functionality. The model defined in this specification shares with HTML the use of URI technology, but goes beyond HTML in offering features, previously available only in dedicated hypermedia systems, that make hyper linking more scalable and flexible. Along with providing linking data structures, Xlink provides a minimal link behavior model; higher-level applications layered on Xlink will often specify alternate or more sophisticated rendering and processing treatments. Till now the Xlink in not implemented inside the browser and the specifications are not very clear. Xlink is not yet fully implemented in the browser and its specifications in very vague.

Hence, the main goal of this article is to develop a comprehensive design of Xlink language.

2 HTML Hyperlinks Limitations

The links supported by HTML are unidirectional and are embedded into the source document. These two factors make it very difficult to locate all documents that link to a given target. We can easily determine all the targets to which a given document is linked by searching the document for <a> elements and extracting the relevant URIs. Achieving the reverse is extremely difficult. Identifying all of the documents that link to a specific target would require retrieving and checking every single page on the Web to determine whether or not it contained a reference to the target.

One of the major consequences of this problem concerns maintainability of the Web. If an author changes or deletes a page, it is very difficult for him to know which other pages may be affected, as it is almost impossible to know which pages may be linked to the one that has been changed. A result of this is that some links are broken and some point to content that is no longer relevant.

Additionally, HTML links have a single-source anchor and a single-destination anchor. This means that it is not possible to create links with multiple destinations. Consider the following scenario: A user navigates to a Web page that contains a description of a research project. The information includes a link to the project researchers, which when followed opens two separate documents containing, respectively, details about the two main researchers. This is not possible to represent using the linking model in HTML, though it is possible to simulate this behavior using other technologies such as scripting languages.

Furthermore, HTML has no mechanism for the typing of links. It is not possible to add attributes of any sort other than a name, a title, and an ID to links. In other words, it is not possible to define different categories of links within an HTML document and then have these categories managed in different ways. This would be particularly useful in supporting more effective browsing and navigation. In many cases it would be useful to be able to show different types of links using different forms of highlighting, as a way of providing additional information to users.

For example, a site might contain structural links that relate to the inherent organization of the content, elaborative links that connect to more detailed information

on a particular topic, and definition links that provide connections into a glossary. Each type of link could be shown differently (different fonts, colors, and so forth).

To overcome all these limitations, an innovative solution is needed and this is the reason to create Xlink.

3 XLINK

The Xlink recommendation describes a XML based linking format. Xlink allows the expression of multi-headed links, out-of-line links, and to associate Meta data with a link. Xlink was originally designed to be the linking standard for XML documents and therefore has some XML specific properties, but the standard does not dictate how Xlink elements should be used.

Linking Functionality

The linking model that has underpinned the traditional Web model is very simplistic. Essentially, the standard <A HREF=...> link is a simple static, directional, single-source, single-destination link that is embedded into the source document.

It is:

- Static because the link never changes.
- Directional because the link has an explicit direction of association.
- Single-source because the link only has one point from which it can be triggered.
- Single-destination because the link only has one resource to which the client traverses when the user activates the link.

A resource can be used in various ways; the most common use in this research will be related to linking. We need to define the relevant regions of the resource that are participating in a link. Within the context of a link, these regions are referred to as anchors or in XML linking, they are referred to as locators.

We also need to distinguish between links and arcs, which in turn requires an understanding of a third concept; traversal. In an HTML “A” link, involves a single source anchor, a single destination anchor, and an implied connection between them. When we display an HTML document and activate the anchor, the link is traversed to the destination link.

Illustrative example:

- The link between a document and an image is traversed automatically when the document is loaded, and the resultant image is embedded into the source.
- The link between a document and a style sheet “specified using a LINK element” is traversed automatically, but does not result in any content being embedded.

In each case, the various link characteristics such as how the traversal of the link is initiated, the behavior upon traversal, and the specification of the link semantics are typically implicit for that link type.

When it comes to linking in XML, the situation is different and more flexible. Xlink allows the definition of links, but a link does not imply traversal. In Xlink a link is simply an association between a numbers of resources specified using locators. There is no link "source", nor is there a link "destination". This is because Xlink has separated the concept of associating resources "using a link" from the concept of traversal amongst these resources. This traversal is specified using arcs. A given link may contain a number of different arcs.

Consider the situation where we have a link that has multiple participating resources, with an arc that starts from more than one resource, and ends on more than one resource. In other words, the arc traversal can be activated from multiple different locations, and when it is activated, it results in the presentation of multiple new resources rather than just a single resource. The result is a multi-source, multi-destination link, something that is not possible to implement using the simple linking model within HTML.

3.1 Simple Web Link

One of the simplest examples of information associations is the ordinary HTML link. Consider the example HTML code shown in Figure 1. This is a simple static, unidirectional, single-source, single-destination link that is embedded into the source document.

- It is static because the link never changes unless it is modified by an author.
- It is a unidirectional because the link has an explicit direction of association and hence, usually, an explicit primary direction of navigation.
- It is single-source because the link has only one point from which it can be triggered. It is single-destination because the link has only one information element that is accessed when the link is activated.
- The link is embedded into the source document (indeed, it is embedded within the source anchor) because the description of the connection between the source and destination anchors exists within the source anchor.

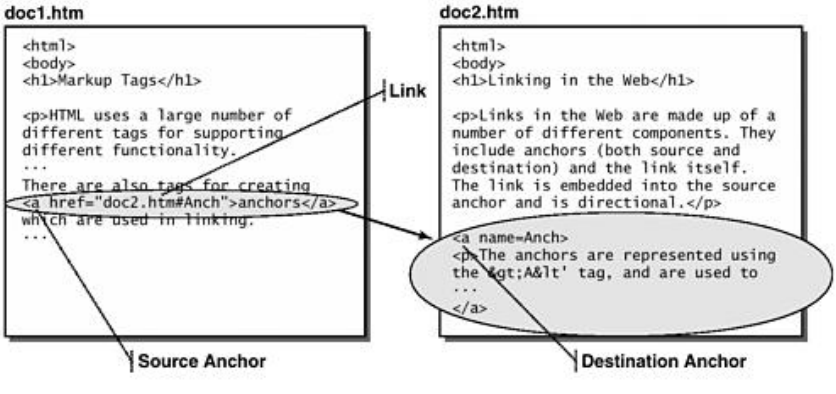


Fig. 1. Simple Web link

3.2 Simple Links in XLink

A simple link defines a one-way link between a starting resource and an ending resource. The starting resource is always an XML element that is defined as a link by the means of XLink syntax. The ending resource is identified by a URI. The URI can, but does not have to be, a URL. A simple XLink link is directly analogous to the standard HTML link that most people are familiar with.

Any XML document that uses XLink must define the XLink namespace. The namespace is "http://www.w3.org/199/xlink", and the namespace prefix associated with it is traditionally "XLink." Although another prefix can be used ,it would serve, I think, only to confuse people. The namespace must be defined at or above the highest (that is closest to the root) element in the document that uses XLink, and it is often done in the root element. For example, in an XML document whose root element is "books," you could define the namespace as such:

```
<books xmlns:xlink=''http://www.w3.org/199/xlink''>
..
</books>
```

Once the namespace is defined, the minimum steps required to create a simple link are to add the following two attributes to the element:

- *xlink:type*, specifies the type of the link. For a simple link, the type is "simple".
- *xlink:href*, specifies the target URI of the link.

The following example creates a link to the indicated URL:

```
<Myelement xlink:type=''simple'' xlink:href=
''http://www.google.com''>
..
</Myelement>
```

It is important to note that the use of the *title* and *role* attributes is entirely up to the processing information. The URI pointed to by the *role* attribute may indicate a file that will be read, or it may just serve as a form of unique identification (similar to the namespace).

A simple link can also contain attributes that define the behavior of the link. The *xlink:show* attribute defines what happens when the link is activated, and the *xlink:actuate* attribute specifies when the link should be activated.

Settings for the *show* attribute:

Value	Description
– embed	– The linked resource will be displayed embedded in the current document.
– new	– The linked resource will be displayed in a new window.
– other	– The linked behavior is defined elsewhere.
– replace	– The linked resource will replace the current document.
– none	– There are no constraints on the link behavior.

Settings for the *actuate* Attribute:

Value	Description
– onLoad	– The link is activated when the document is loaded.
– onRequest	– The link is activated when some event occurs. For example, when the cursor clicks the link.
– other	– The link activation behavior is defined elsewhere.
– replace	– There are no constraints on the link activation

3.3 Dynamic Links

We can also have dynamic links, where the destinations resources are only resolved when traversal occurs. A dynamic link will have as structure or a behavior that change over time, or with different users or circumstances. These changes can be with the link structure "such as changes in link destinations" or with the link semantics "such as whether the activation results in the new resource replacing the existing content, or being embedded into the existing content". The most common example of dynamic links are where the link destination is a service rather than a document, such as a CGI script. These types of links can be important for supporting adaptive systems.

Like the previous case, this is directional, single-source, single-destination link that is embedded into the source document. The major difference is that the link is dynamic rather than static. The link is a computer-generated interface "CGI" that analyzes relevant input data and determines a suitable document to be returned (either directly or through an appropriate HTTP redirect response). For example, the CGI program may return a time-dependent page based on the current time or a user-specific page based on user's previous navigation "as stored in cookies". In other words, the destination may change each time the link is followed. Note also that the link is not multi-destination, despite the existence of multiple possible destinations. Each time the link is activated, only a single destination is retrieved.

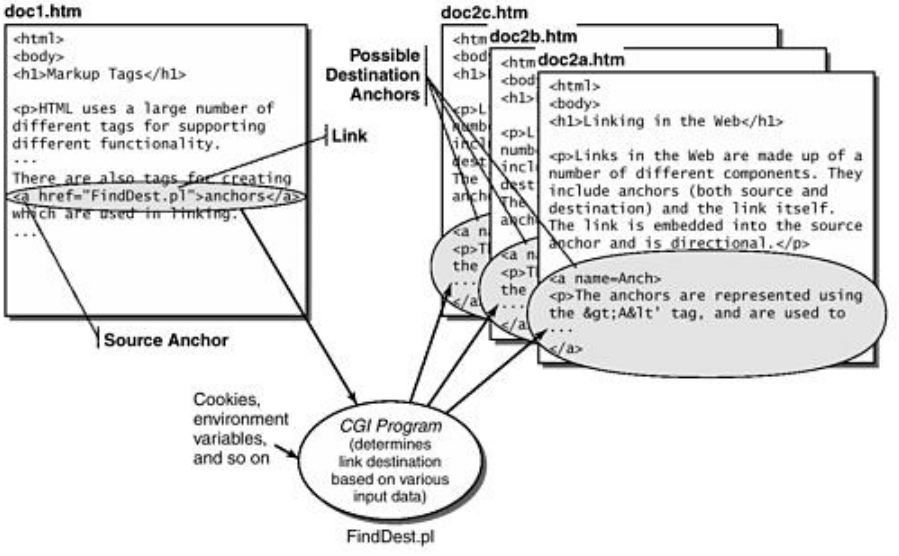


Fig. 2. Dynamic Web link

Although HTML has no support for dynamic links, it is possible to create them using server-side technologies such as CGI scripts, possibly in conjunction with other technologies such as HTTP cookies or URL rewriting.

3.4 Extended Links in Xlink

An extended link describes a collection of resources and the paths along them. An individual resource in the collection can be linked to any or all of the other resources, to none of them and even back to itself. In effect, an extended link defines relationships between resources. Extended links describe connections between any number of resources in any combination of local or remote. The link says nothing about what is actually done with these relationships that is up to the processing software. At present, software support of extended links is essentially nonexistent, so the types of uses that will be devised for extended links remain to be seen.

To create an extended link, use the *xlink:type* attribute with the value “extended”. The following example defines the “filmreport” element as an extended link:

```
<filmreport xlink:type=''extended''>
..
</filmreport>
```

An *extended* link can have *xlink:title* and *xlink:role* attributes that serve the same purpose as they do in the *simple* link. Other link details are provided by elements that are placed within the link element. These elements are as follows:

- *locator*: Identifies a remote resource
- *resource*: Defines a local resource
- *arc*: Defines traversal information between two resources
- *title*: Defines a title for the link

The Locator Element
Locator-type elements are used to link to remote resources. The *locator* element identifies an individual remote resource that is part of the *extended* link. A *locator* element has five attributes associated with it.

- *xlink:type*: Required. Must be set to "locator".
- *xlink:ref*: Required. The URL of the remote resource.
- *xlink:role*: Optional. A URI that provides information about the resource being pointed to by the link.
- *xlink:title*: Optional. The title of the link.
- *xlink:label*: Optional. A name identifying the resource. A label must be provided if the link is to be referenced in an arc element.

The *xlink:role* an *xlink:title* attributes play the same role here as they do with a *simple* link.

The Resource Element
In *extended* links *resource*-type elements are used to link local resources. The information provided by a local resource is contained in the element itself and is not external to the XML file. A *resource* element must have the *xlink:type* attribute set to "resource". It can also have optional *xlink:title*, *xlink:role*, and *xlink:label* attributes that serve the same purpose as described previously for the *locator* element. In almost all cases, a *resource* element will contain child elements that hold the actual information.

The Arc Element
In *simple* links, the *show* and *actuate* attributes are used to determine how and when links will be traversed. In *extended* links, the *show* and *actuate* attributes are used with *arc*-type elements to specify traversal information between participating resources.

Extended links can create more kinds of links than *simple* links can. Therefore, additional attributes are necessary to specify the start and end points of the link. *arc*-type elements inside of *extended* links use attributes called *xlink:from* and *xlink:to* to specify the starting and ending points of links.

Each *arc* has only one starting resource and one ending resource. The values of the *xlink:from* and *xlink:to* attributes must be the values of *xlink:label* attributes from other linking elements inside of an *extended* link.

The *arc* element specifies traversal information among an *extended* link's participating resources. In order to use a resource in an *arc* element, the *resource* or *locator* element must be assigned a name using the *xlink:label* attribute.

The arc element has the following attributes:

- *xlink:type*: Required. Must be set to "arc".
- *xlink:title*: Optional. The title of the arc.
- *xlink:arcrole*: Optional. A URI that provides information about the link (similar to the *xlink:role* attribute for other elements).
- *xlink:show*: Optional. Specifies what happens when the link is activated.
- *xlink:actuate*: Optional. Specifies what happens when the link is activated.
- *xlink:from*: Optional. Identifies the from resource by its *xlink:label* attribute. If omitted, all resources in the extended link are treated as sources.
- *xlink:to*: Optional. Identifies the "to" resource by its *xlink:label* attribute. If omitted, all resources in the extended link are treated as targets.

The Title Element

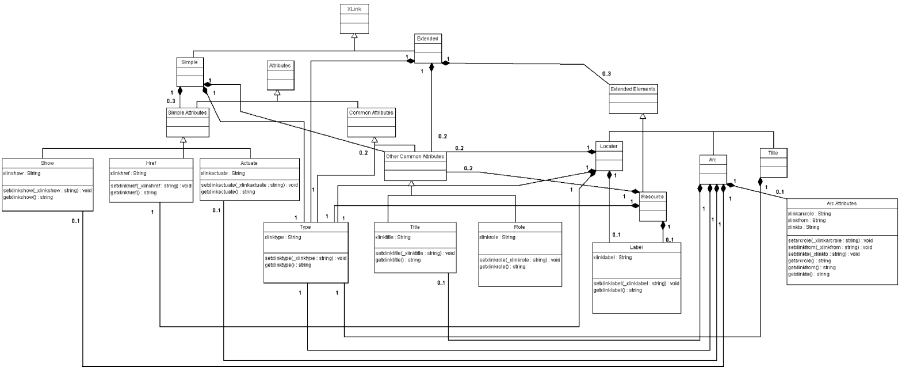
The *title* element can appear as a child of extended, locator, resource, or arc elements, and it can contain its own child elements. Using the title element provides more flexibility than using *the xlink:title* attribute.

After this detailed description of Xlink specification, we will present our new design of Xlink using UML class diagram.

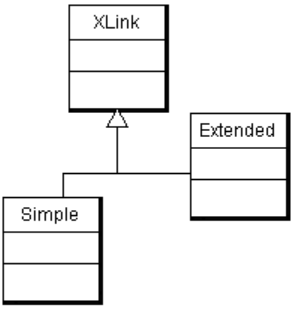
4 New Design of XLINK

In this section, we propose a new, clear and practical design for Xlink using UML [15] class diagram. We choose UML class diagram to be easily implemented it.

Overall Diagram

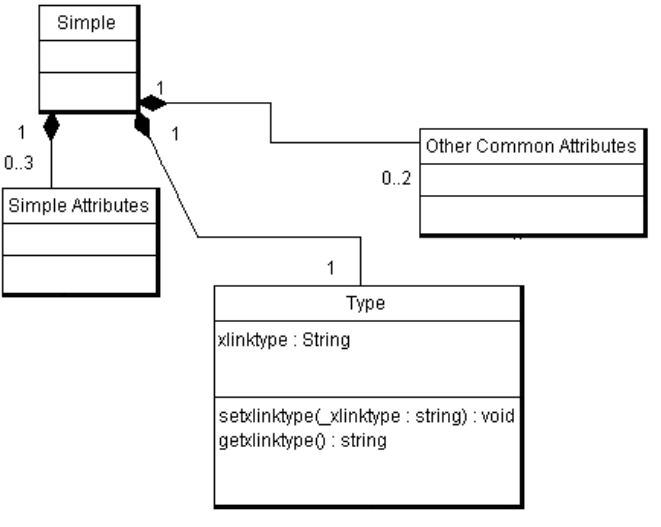


XLink Class



In the picture on the right we have a Super Class called “XLink” and two sub classes named “Simple” and “Extended”. The “Simple” class represents the class for simple xlinks and the “Extended” class represents the extended xlinks.

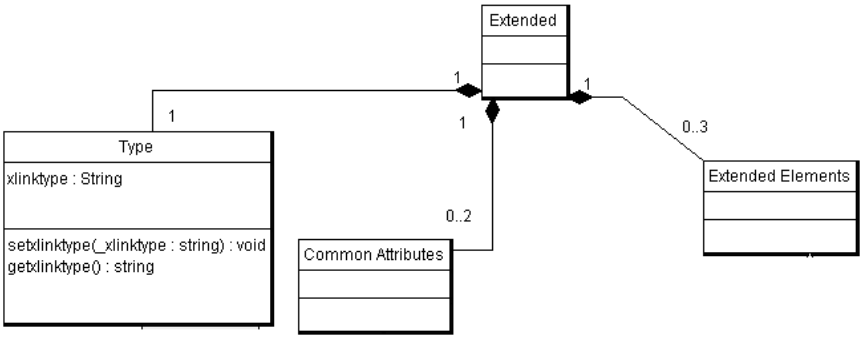
Simple Class



The “Simple” class is composed of “Simple Attributes”, that are attributes found in simple Xlink and not in extended Xlink. The “Simple” class can be composed of a min 0 and max 3 “Simple Attributes”.

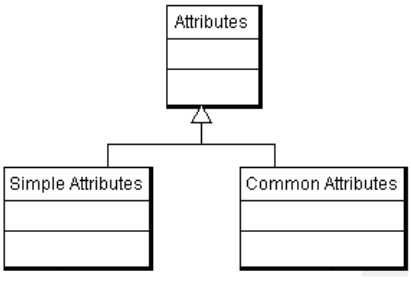
"Simple" class is also composed of one "Type" that defines the type of the Xlink and in this case it must be simple and it is required to be present in a simple Xlink. "Other Common Attributes" are attributes that are found in simple Xlink and extended Xlink as we will see later. It can be composed of a min 0 "Common Attributes" and a max of 2.

Extended Class



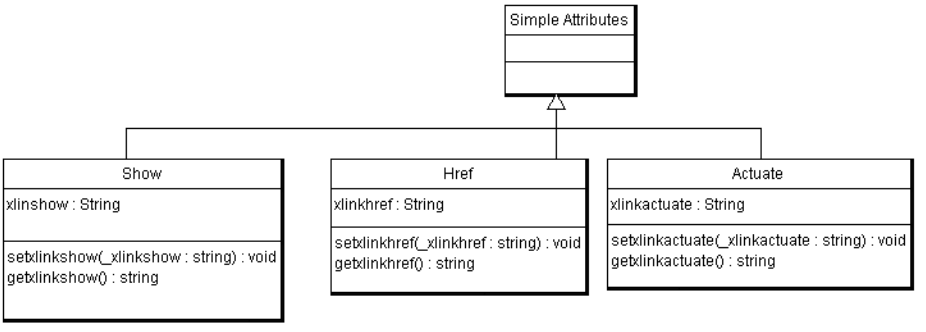
The "Extended" Class is composed of one "Type" attribute which is a required attribute, and the "Common Attributes" mentioned before. It can have a min of 0 "Common Attributes" and a max of 2. It is also composed of "Extended Elements" that are sub elements of the original extended element written in XML. It is composed of a min 0 and a max of 3 of "Extended Elements".

Attributes Class



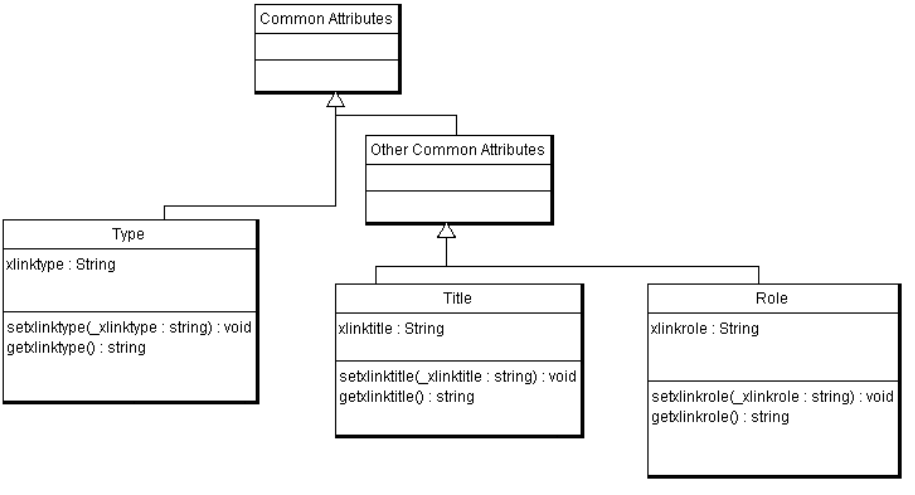
Since "Simple Attributes" that are for the simple Xlink and the "Common Attributes" are both attributes, we create a "Attributes" class and we make inheritance from it to both "Simple Attributes" class and "Common Attributes" class.

Simple Attributes Class



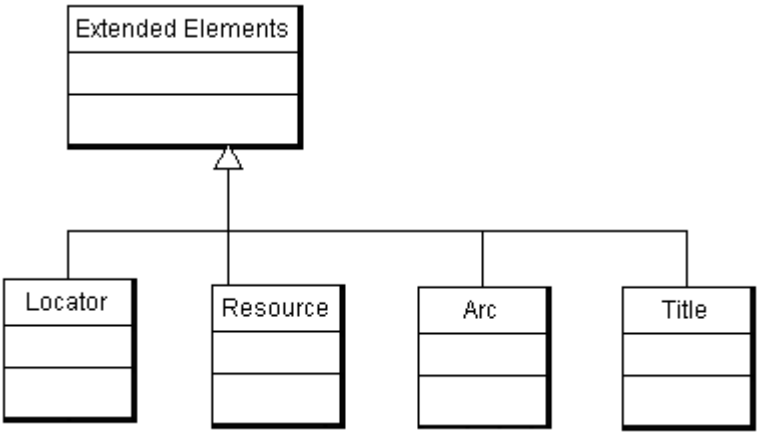
"Show", "Href" and "Actuate" are all simple attributes, i.e. are attributes used to describe a simple Xlink.

Common Attributes Class



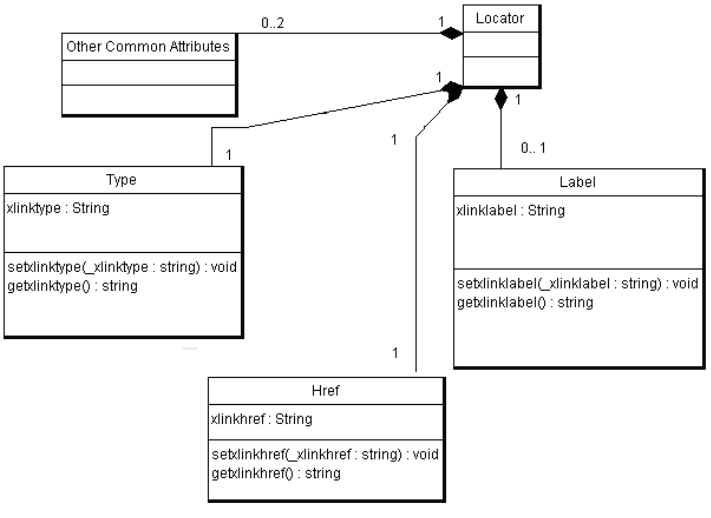
We have separated "Type" class from "Title" and "Role" because in some cases other classes are composed of "Type". But as we see "Type", "Title" and "Role" have inheritance from "Common Attributes".

Extended Elements



The classes “Locator”, “Resource”, “Arc” and “Title” are sub classes of “Extended Elements”. These four are represent elements in the xml document and are sub elements of the parent element holding the extended Xlink attribute.

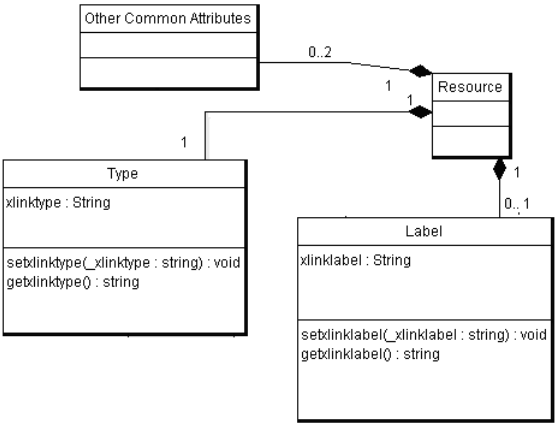
Locator Class



The “Locator” class represents the element locator in an xml file, which in fact is a sub element of the element holding the extended Xlink. A “Locator” element in the xml file is composed of several attributes, “Type”, “Href”, “Label” and “Other Common Attributes” which were explained earlier. It must have a type attribute. It

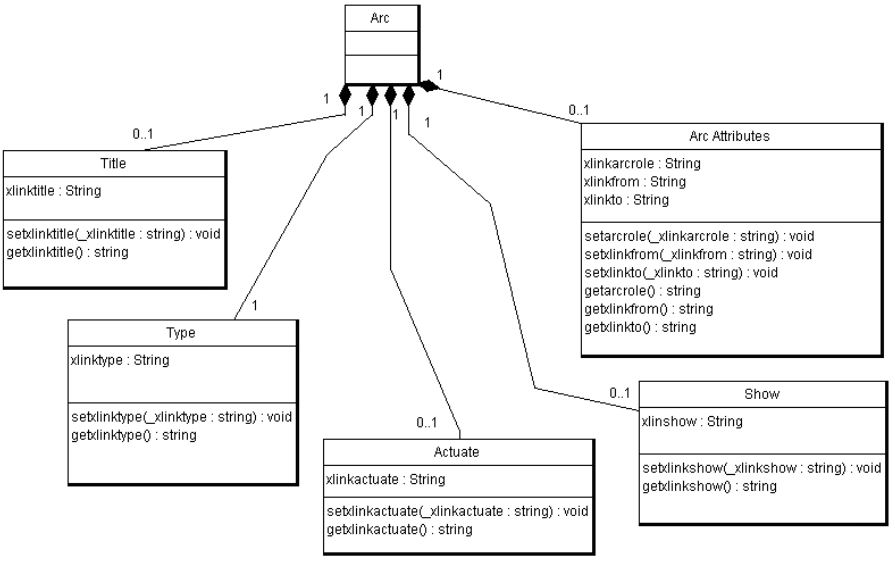
must have one “href” and one type attributes. In addition it can have a min of 0 and a max of 2 other common attributes, it can have a label attribute.

Resource Class



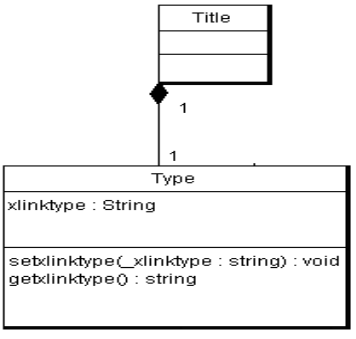
The “Resource” class is composed of a min of 0 other common attributes, it can have only one label attribute and must have a single type attribute.

Arc Class



The “Arc” class represents a sub element of a parent element that has a type of extended Xlink. It must have a type attribute, it can have a title, actuate, show attributes. It also has the option of having the three attributes found in “Arc Attributes”, all of them or anyone of them or none.

Title Class



The “Title” class has only one attribute which it must have and it is called type.

5 Conclusion

A clear and practical design of Xlink is more than needed. In this article, we propose a new design for Xlink language using UML class diagram. Our future work is to implement this design inside the web browser.

Acknowledgments. This work was funded by the Lebanese university.

References

1. Aitken, P.G.: XML – The Microsoft Way. Addison-Wesley (2002)
2. Valentine, C., Minnick, C.: XHTML. New Riders (2001)
3. Lowe, D.: Improving Web Linking using Xlink, University of Technology Sydney (2001)
4. http://www.xml.com
5. http://www.w3schools.com
6. http://www.w3.org
7. http://xml.coverpages.org
8. Wilde, E., Lowe, D.: Xpath, Xlink, XPointer, and XML: a practical guide to Web hyper linking and transclusion. Addison-Wesley Professional (2003)
9. Harold, E.R., Scott Means, W.: XML in a nutshell. O’Reilly Media, Inc. (2004)
10. Harold, E.R.: XML 1.1 bible. Wiley Pub. (2004)
11. Ray, E.T.: Learning XML. O’Reilly Media, Inc. (2003)

12. Shelly, G.B., Woods, D.M., Dorin, W.J.: HTML: Comprehensive Concepts and Techniques. Cengage (2008)
13. Kadry, S., Smaili, M., Kassem, H., Hayek, H.: A New Technique to Backup and Restore DBMS using XML and .NET Technologies. International Journal on Computer Science and Engineering (IJCSE) 2(4), 1092–1102 (2010)
14. Kadry, S., Kassem, H.: Design of Heterogeneous Databases Replication Using XML. International Journal of Reviews in Computing 1 (December 2009)
15. Miles, R., Hamilton, K.: Learning UML 2.0. O'Reilly Media, Inc. (2006)

Design and Implementation of Ubiquitous Pig Farm Management System Using iOS Based Smart Phone

Jeong-hwan Hwang and Hyun Yoe

School of Information and Communication Engineering,
Sunchon National University, Korea
{jhwang,yhyun}@sunchon.ac.kr

Abstract. The interest in smart phone has been increasing in recent from the advancement of ubiquitous and mobile technology and smart phone is bringing forth significant changes in our lives by allowing us to obtain and share information anywhere, anytime by connecting to the Internet. The purpose of this paper is to propose a system for managing pig farm anywhere, anytime via smart phone to ultimately establish ubiquitous farming environment and improve livestock farms productivity. In this system, WSN environmental sensors and CCTVs are installed in pig farm that will collect data on livestock breeding environment such as illumination, humidity, temperature and gas to allow gathering and monitoring of pig farm environment. This system not only allows users to control and monitor pig farm facilities remotely via smart phone but also helps to create optimal breeding environment through the breeding environment data obtained over a long period of time. For the purpose of verifying this ubiquitous pig farm management system using smart phone, a pig farm model was developed and tested by applying the system.

Keywords: WSN, Pig, Ubiquitous, Agriculture, Smart phone.

1 Introduction

The mobile phone that used to be used primarily for phone conversation is now becoming an application-centered mobile Internet device at the recent advancement of ubiquitous and mobile technology, and people are now able to obtain and share information anywhere, anytime upon connecting to the Internet[1][2].

The advent of smart phone is bringing forth significant changes not only in the field of IT but also in our lives and many studies are being conducted on the effective use of smart phone[3][4].

The Korean livestock industry is currently facing a situation where it must go face-to-face with countries with advanced livestock industries as a result of recent feed price increase and signing of FTA(Free Trade Agreement). In addition, many hog breeders are experiencing significant difficulties that are caused by the increase in

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 147–155, 2011.

livestock death rate from various consumptive diseases, as well as feed price, raw and subsidiary material price and energy cost[5][6].

Livestock breeders need to create optimal breeding environment by utilizing systematic and scientific breeding technologies to cope with and solve the issues they are facing, as well as increase productivity and produce high-quality stock farm products by reducing livestock death rate and production cost[7][8].

In this paper, a ubiquitous pig farm management system will be proposed that can monitor pig farm environment and control building facilities via smart phone to solve the issues Korean livestock industry is currently facing.

In this system, sensor nodes are installed in pig farm to establish a WSN and measure pig farm environment to collect data. In addition, CCTVs are used to collect video data on pig farm. The pig farm environment and video data collected are used to monitor the pig farm anywhere, anytime via smart phone and this system can provide user convenience and increase productivity by allowing users to control their pig farm facilities.

This paper consists as follows: the structure of the proposed ubiquitous pig farm management system using smart phone will be explained in Chapter 2, and testbed for the system will be developed and the performance will be evaluated in Chapter 3 and the thesis will be concluded in Chapter 4.

2 Design of the Ubiquitous Pig Farm Management System

2.1 System Structure

The proposed ubiquitous pig farm management system using smart phone can be classified into three layers, as shown in Figure 1, and the three layers include physical layer that consists of environmental sensor, CCTVs and pig farm control facilities, application layer that consists of the interface that supports pig farm environment monitoring and control service, and middle layer that supports the communication between the physical and application layers, converts pig farm data into database, provides monitoring & control service and maintains optimal hog breeding environment.

The physical layer consists of environmental sensors for collecting pig farm environment data, CCTVs for collecting pig farm video data and pig farm facilities for creating optimal stock breeding environment. The environmental sensors are installed in pig farm to collect stock breeding environment data such as illumination, humidity, temperature and gas, and the sensor nodes, upon autonomously creating a network, wirelessly collects physical data obtained from the sensor nodes and measure environment changes.

CCTVs are installed both inside and outside of pig farm and the indoor CCTVs are installed to collect pig farm and hog video data and the outdoor CCTVs are installed to prevent theft, fire, etc.

The pig farm facilities refer to the devices that control the environmental elements that affect livestock growth such as illumination, temperature, humidity and gas, and they include lighting device, humidifier, air conditioner, ventilation fan, etc.

The middle layer consists of sensor manager for managing environment data collected from the physical layer sensors, video data manager for managing video data collected from CCTVs, pig farm facility manager for managing pig farm facilities, pig farm database that stores pig farm data and pig farm management server for monitoring pig farm and controlling its facilities.

The sensor manager stores into pig farm database the pig farm environment data collected from the physical layer environmental sensors by formatting them into storable format, converting them into units according to measurement elements and using update inquiry for the processed data.

The pig farm facility manager, upon receiving control signal, operates/manages pig farm facilities, and plays the role of storing the status of pig farm facilities into pig farm database. The video data manager provides stream data to the web.

The pig farm database stores in respective tables the pig farm environment data such as illumination, temperature, humidity and gas collected from the sensors installed inside and outside of pig farm, in addition to storing the pig farm video data collected from CCTVs, pig farm facility status & control data and the environment standard values for auto-control & status notification.

The pig farm management server is located between the user and the pig farm database, and periodically tests and notifies the user the environment data stored in the pig farm database, and compares them with the environment standard values stores in the pig farm facility control table to control the facilities.

The application layer consists of applicant services that support various platforms such as laptop, web, PDA and smartphone, and it provides users with pig farm environment monitoring service, pig farm video monitoring service and pig farm facility control service.

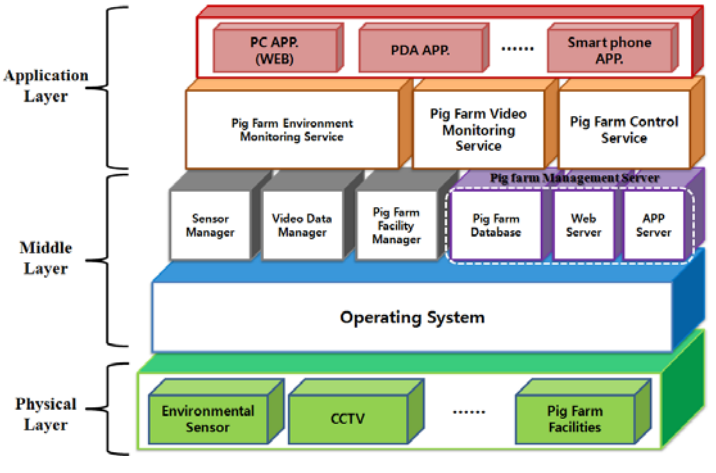


Fig. 1. Ubiquitous Pig Farm Management System Structure

2.2 Service Process

One of the major services provided by the proposed system is the pig farm facility control service for allowing the pig farm management server to automatically control the pig farm facilities or allowing the breeder to manually control them based on the data collected through the environmental sensors and CCTVs installed both inside and outside of pig farm.

The automatic control service saves the information collected from pig farm at pig farm database. The pig farm management server calls up the information and compares it with the environmental standard values saved in the pig farm database. If it is more than or short of standard value, it will confirm whether the pig farm facilities are operating as saved in the pig farm database. Then it will send the control signal to pig farm facilities manager and control the pig farm facilities.

When pig farm facilities operate, the pig farm facilities status information is saved in the pig farm database and it will be notified to user. Figure 2 shows the operation process of pig farm facilities automatic control service.

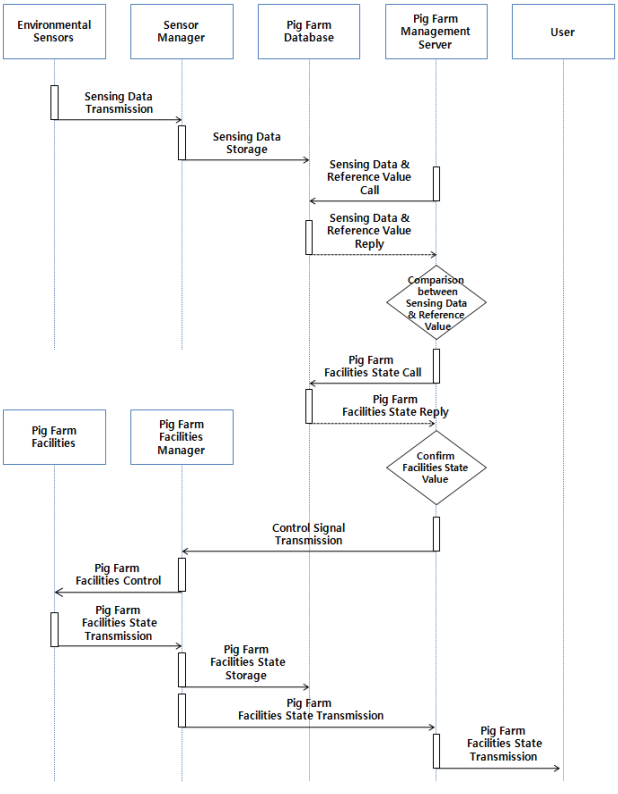


Fig. 2. Service Process of the Pig Farm Facilities Automatic Control

The manual control service saves the information collected from pig farm in the pig farm database and the pig farm management server sends the information to the user in real time. If the user wants to control the pig farm at this time, the user will send the pig farm facilities control signal to pig farm management server through GUI. The pig farm management server will check whether the pig farm facilities are operating through pig farm database and send the control signal to pig farm facilities manager to control the pig farm facilities. Figure 3 shows the operation process of pig farm facilities manual control service.

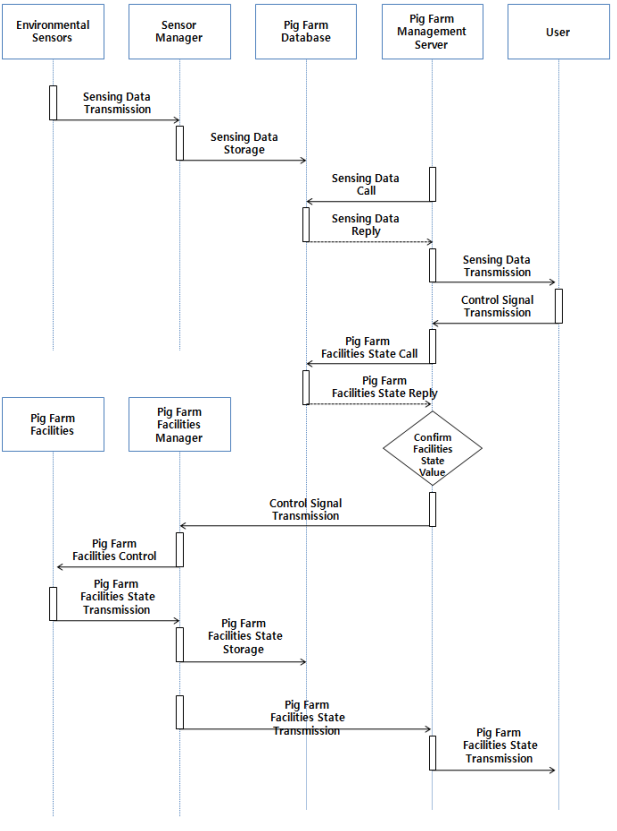


Fig. 3. Service Process of the Pig Farm Facilities Manual Control

3 Implementation of the Ubiquitous Pig Farm Management System

For the purpose of evaluating the performance of the proposed system, a pig farm model has been created, as shown in Figure 4, and applied the proposed system to establish a ubiquitous pig farm management system.

Fig. 4. Ubiquitous Pig Farm Model & Control PC

This ubiquitous pig farm management system refers to a system that can monitor the pig farm environment and control the pig farm environment control facilities by applying WSN technology to pig farm.

In the pig farm model, environmental sensors were installed for pig farm environment monitoring and web cam was also installed for video monitoring of the pig farm. In addition, environment control facilities such as lights, heating fan and ventilation fan to maintain optimal hog breeding environment, and ubiquitous pig farm management system web GUI, as shown in Figure 5, was developed to control the system and conduct monitoring test.

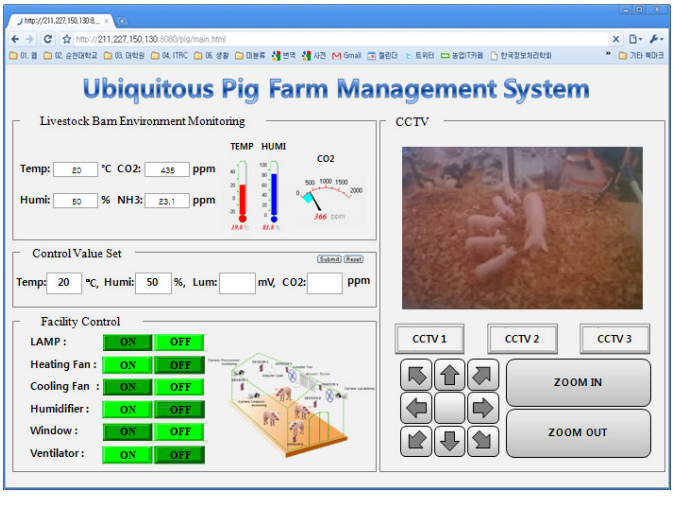


Fig. 5. Ubiquitous Pig Farm Management System Web GUI

With regards to the web development environment, the PC that works as server was set according to general PC environment and the sensor node environment for sensing was set according to environment ideal for using Zigbee sensor. In addition, Tomcat-6.0.20 was used for the WAS and MySQL v.5.0 that is the most stable version available was used for the database.

Table 1. Hardware Environment

	Type	Details
Server PC Environment	CPU	Intel Xeon 3.2 Ghz
	RAM	1 GB
	OS	Microsoft Windows XP

Table 2. Software Environment

	Type	Details
PC Development Environment	OS	Microsoft Windows XP
	Programming Languages	JAVA (JDK 6), C#
	RDBMS	MySQL 5.0
Smart Phone Application Development Environment	IDE	Xcode 3
	Programming Languages	Objectives-C
	OS	MAC OS X 10.6.7
	Mobile OS	IOS 4.3.3
Sensor Node Environment	OS	TinyOS 1.0
	Linux Environment	Cygwin
	Programming Languages	NesC
	JAVA	JDK 1.4.1

As shown in Figure 6, the environment and video data of pig farm can be checked in real-time through iOS-based iPad application, in addition to controlling the pig farm facilities.

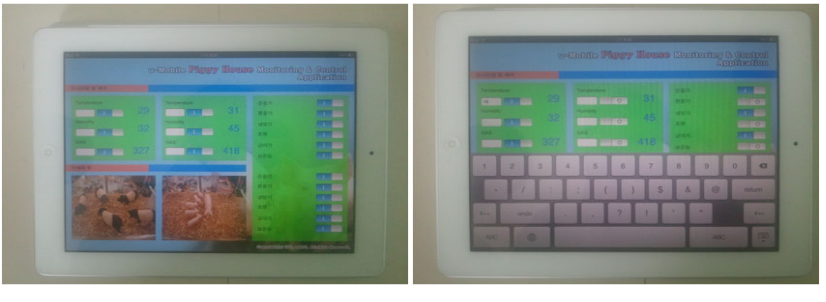


Fig. 6. iPad Application GUI of Ubiquitous Pig Farm Management System

The pig farm facilities can be controlled automatically or manually. In the case of auto-control, the user enters pig farm environment standard values by using the iPad

application, which are stored in the pig farm database. The values are compared with the environment data collected through sensors in real-time and the pig farm facilities are controlled accordingly when the collected data surpass or fall short of the environment standard values.

The result of testing the environment data monitoring and facility controlling of the pig farm model by using the iPad application revealed that the system is operating smoothly.

4 Conclusions

In this paper, a ubiquitous pig farm management system using smart phone was proposed to ensure systematic and scientific livestock breeding technology.

The proposed system consists of physical layer for collecting pig farm data and controlling the environment, and application layer that consists of interface that supports service, and middle layer that supports the communication between the physical and application layers, and converts pig farm data into database, and provides monitoring & control service, and maintains optimal hog breeding environment.

A pig farm model was created and the proposed system was applied to test the system, and the test result showed that the proposed system helped to maintain optimal stock breeding environment and provided user convenience.

It is expected, accordingly, that applying the proposed system to stock breeding farmhouses will help them reduce labor cost, produce high-quality products and ensure competitiveness.

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-2011-(C1090-1121-0009)).

This research was financially supported by the Ministry of Education, Science Technology (MEST) and National Research Foundation of Korea(NRF) through the Human Resource Training Project for Regional Innovation.

References

1. Lee, S.Y., Lee, H.K., Kim, W.S., Lee, J.H., Kim, S.J.: The Trend of Smartphone Operating System Development. Electronic Communications Trend Analysis, Korea Electronics and Telecommunications Research Institute 19(6) (2004)
2. Kim, D.H., Ryu, C., Lee, J.H., Kim, S.J.: Mobile Software Platform Trends for Smartphone. Electronic Communications Trend Analysis, Korea Electronics and Telecommunications Research Institute 25(3) (2010)
3. Kim, Y.-B.: Presence and Future of the SmartPhone and Security Apps. Journal of Korea Contents Association 8(3), 42–74 (2010)
4. Kim, J.-S.: Technology Introduction and Revitalization Plans of Smart Phone. Journal of Korea Contents Association 8(2), 34–38 (2010)

5. Lee, J.-H.: (Livestock industry research series 11) What is a threat to South Korea's livestock industry? Focus attention GSnJ No. 55 (2008)
6. 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)
7. Hwang, J.H., Yoe, H.: Study of the Ubiquitous Hog Farm System Using Wireless Sensor Networks for Environmental Monitoring and Facilities Control. Sensors 10, 10752–10777 (2010)
8. Lee, S.-C., Kwon, H., Kim, H.-C., Kwak, H.-Y.: A design and implementation of Traceability System of black pigs using RFID. J. Korea C. Assoc. 8, 33–40 (2008)

Design of Cattle Barn Management System Based on Thermal Imaging Data

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. This paper is a cattle barn management system based on thermal imaging data. This system is for recognizing the condition of livestock and the changes in environment by utilizing thermal imaging data and sensor data, as well as for automatically controlling the optimal growth environment and diagnosing early livestock diseases by analyzing and utilizing the data collected. Accordingly, the system has been composed into Physical layer, middleware, cattle barn control part and web-GUI. The thermal imaging data & Physical layer collects in realtime livestock condition data and cattle barn data, and the middleware manages the collected data. In addition, the cattle barn control part manages with intelligence any occurrence of livestock and cattle barn abnormalities. Lastly, the system provides the service of allowing user to remotely monitor cattle barn environment through the web-GUI.

Keywords: Cattle barn, thermal, livestock.

1 Introduction

In recent, livestock contagious diseases have broke out in domestically such as foot-and-mouth disease and Avian Influenza (AI). In the case of foot-and-mouth disease in particular, the amount of damage estimated by the government is about 3 trillion won and the amount of damage caused by livestock contagious disease for the past four years since 2006 is nearly seven times of 450.3 billion won. Additionally, the damage resulting from the spread of Highly Pathogenic Avian Influenze (HPAI) is also enormously high [1]. Since there are no solutions for this currently in the country, enormous amounts of damages are being incurred. In addition, the result of comparing the agricultural and livestock industry technological know-how level both domestically and internally shows the following facts: our technological know-how for food crops or high-quality stability production is relatively close to that of advance nations, but our technological know-how for eco-friendly & safe agricultural and livestock farm product production and mechanization & automation significantly fall behind by more than five years from that of advanced nations. In addition, it is being revealed in the livestock breeding productivity, which shows the meat production amount from breeding per head, that our productivity is higher than that of France or

* Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 156–160, 2011.

Australia but 34% lower than that of the US, and that our pig and chicken productivity is lower than that of advanced nations [2]. Accordingly, this paper will propose a cattle barn management system based on thermal imaging data to increase the productivity in the area of livestock breeding, which is labor intensive and has an industry structure that lags behind, and systematically manage livestock diseases and verify the safety of stock farm products. The system that is being proposed has been designed to collect and save data on cattle barn environment, livestock breeding and diseases from the thermal imaging camera and sensor installed within cattle barn, and create situation data to accurately control the cattle barn. This paper consists of three parts - introduction, main subject and conclusion. The specific functions of the system are explained in the main subject part followed by conclusion.

2 Design of Cattle Barn Management System Based on Thermal Imaging Data

The thermal imaging data-based livestock disease diagnosis system consists of thermal imaging data & Physical Layer, middleware, Cattlebarn control and web-GUI.

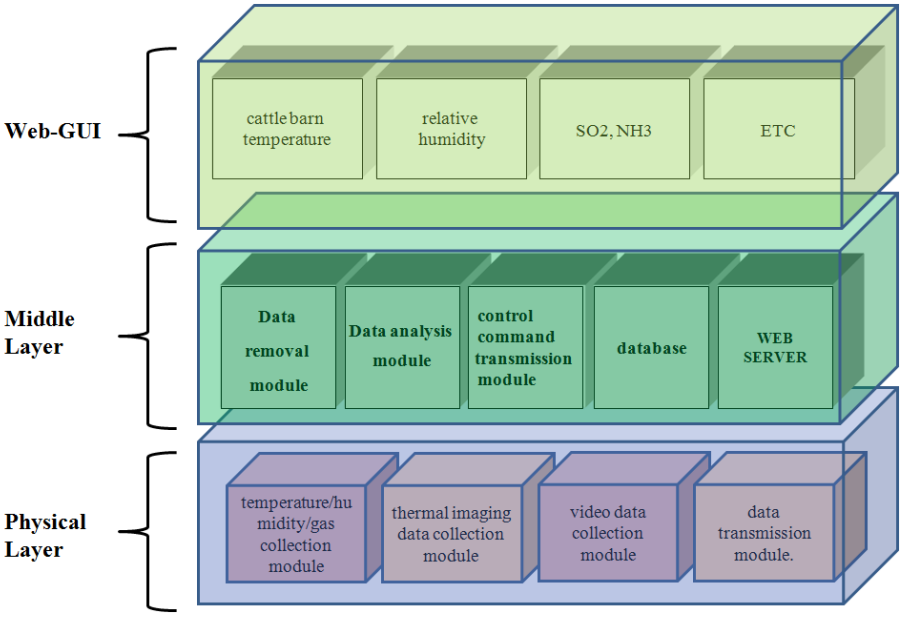


Fig. 1. System Structure

2.1 Physical Layer

As shown in Figure 2, Physical Layer consists of temperature/humidity/gas collection module, thermal imaging data collection module, video data collection module and data transmission module. The temperature/humidity/gas collection module collects temperature/humidity data that are most sensitive to livestock health and measures the level of odor occurring at cattle barn through gas sensor. The thermal imaging data collection module collects livestock condition data such as their temperatures and any diseases through the thermal imaging camera installed within cattle barn. The video data collection module collects video data to allow monitoring of cattle barn environment remotely. The data transmission module transmits the raw data collected to the middleware.

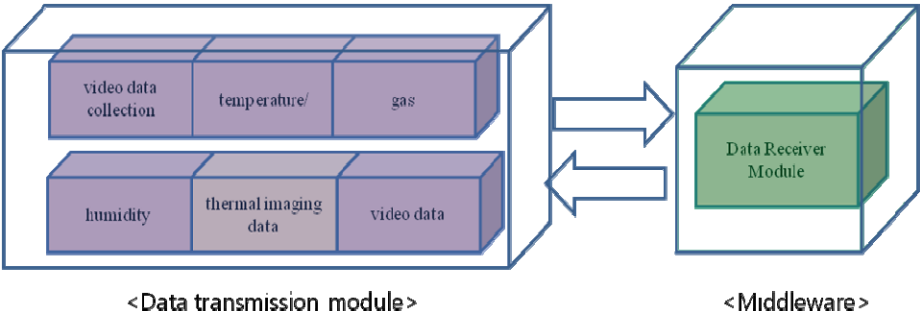


Fig. 2. Physical Layer Structure

2.2 Middleware

As shown in Figure 3, the middle consists of duplicate data removal module, data analysis module, control command transmission module, database and web-GUI data transmission module. The duplicate data removal module removes numerous duplicate data coming in through the data transmission module of Physical layer. The data analysis module analyzes data such as optimal temperature/humidity and ammonia value that are needed to breed livestock based on the data collected from Physical layer, and transmits the data to the database and the control command transmission module. The control command transmission module transmits to the cattle barn control management part according to the optimal data value analyzed through the data analysis module to allow intelligent control of cattle barn. Lastly, the web-GUI data transmission module transmits the analyzed values to the web server through the physical layert and the analysis module.

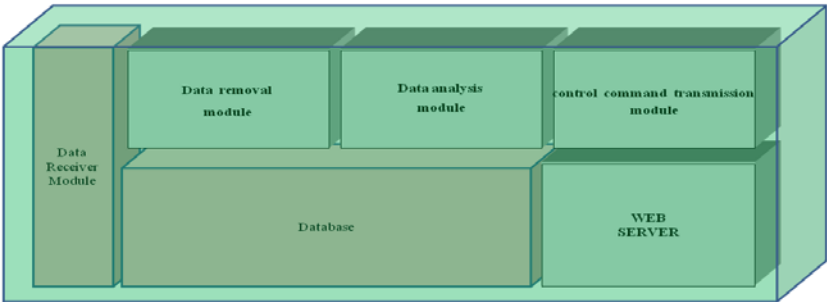


Fig. 3. Middleware Structure

2.3 Cattle Barn Control

The cattle barn control part consists of cattle barn facility devices such as lighting equipment, ventilation fan, fan/fan heater and feeder, as well as PLC that controls these devices. In addition, the cattle barn control part identifies and transmits to the server the operational status of each device and controls the facility devices.

2.4 Web-GUI

The web-GUI through the web-interface provides the cattle barn temperature, relative humidity, livestock breeding luminosity, air flow rate and light penetration amount, as well as the feed amount, remaining feed amount, water supply amount, indoor temperature and indoor harmful gas level (SO2, NH3, etc.) that affect the cattle barn livestock breeding. It also provides the vital data of livestock and disease data through the thermal imaging camera to anywhere with Internet accessibility. In addition, it allows user to remotely monitor the cattle barn through the video data collected at the Physical layer and allows the controlling of facilities from remote places for any occurrence of abnormalities at the cattle barn.

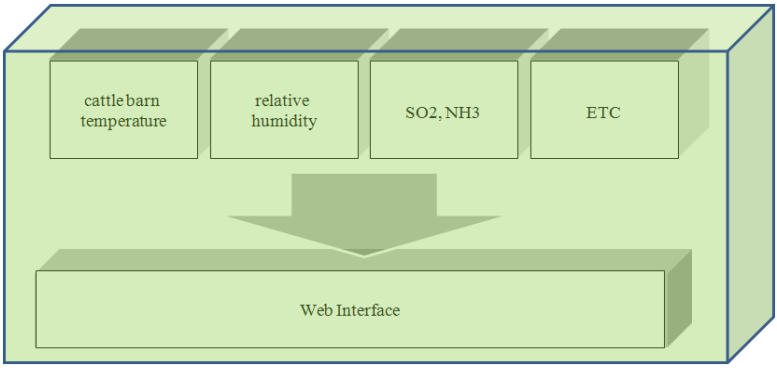


Fig. 4. Web GUI structure

3 Conclusion

In this paper, a thermal imaging data-based livestock disease diagnosis system was designed. This livestock disease diagnosis system has been designed to allow realtime collection of cattle barn environment data and vital data on livestock by utilizing thermal imaging data and sensor data, and to allow the intelligent control of the cattle barn. It is expected that this system will provide significant benefits to farms by reducing the livestock death rate and increasing the production amount. As for future research assignment, a context-based middleware will be developed and applied.

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-2011-(C1090-1121-0009)).

This research was financially supported by the Ministry of Education, Science Technology (MEST) and National Research Foundation of Korea(NRF) through the Human Resource Training Project for Regional Innovation.

References

1. Hong, S.-G.: A Concept Model for Digitalized Animal Disease Control System. Department of Knowledge Management Graduate School of Information Science Soongsil University (2009)
2. Ministry for Food, Agriculture, Forestry and Fisheries; Food, Agriculture, Forestry and Fisheries Statistical Year Book (2008)

Design and Implementation of Wireless Sensor Network Based Livestock Activity Monitoring System

Jeong-hwan Hwang and Hyun Yoe*

School of Information and Communication Engineering,
Sunchon National University, Korea
{jhwang,yhyun}@sunchon.ac.kr

Abstract. Recently in Korea, for the damage caused by the outbreak of livestock diseases such as foot-and-mouth disease and AI has been serious, in order to reduce the damage caused by such livestock diseases, it is necessary to develop collection and analysis technology of livestock biometric data which enables early diagnosis of livestock diseases. In this paper, it is intended to propose the system which is able to collect and monitor livestock activity data by using WSN technology. The system being proposed measures livestock activity by attaching the sensor node on the livestock, compares the collected livestock activity data with the data of activity changes by livestock diseases stored in a database and informs the producer for prompt actions if it is over or under the reference values. Livestock farmers will be able to increase production efficiency and productivity and to minimize damage from livestock diseases by early diagnosis of livestock diseases through biometric monitoring of livestock.

Keywords: WSN, Livestock, Ubiquitous, Agriculture, Activity.

1 Introduction

Korean livestock industry has to directly confront with advanced livestock countries due to recent increase in feeds price and FTA(Free Trade Agreement) agreements, as well as a lot of livestock farmers are facing difficulties from increase of production costs such as feeds, raw materials and energy[1][2][3].

In addition, the damage caused by the recent outbreak of livestock diseases such as foot-and-mouth disease and AI(Avian Influenza) has been serious. For foot-and-mouth disease, the damages have reached up to 1.6 million dollar, and at the same time, for AI has spreaded out, the damages have been enormous[4][5].

However, currently in Korea, there is no system for surveillance and early detection of livestock diseases. In order to minimize damage from such livestock diseases, it is necessary to develop collection and analysis technology of livestock biometric data which enables early diagnosis of livestock diseases.

* Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 161–168, 2011.

In this paper, it is intended to propose the system which is able to collect and monitor livestock activity data by using WSN(Wireless Sensor Network) technology.

WSN technology means the technology utilized for surveillance and control as physical data is wirelessly collected from sensor nodes with computing capability and wireless communication capability which are placed in various application environment and then autonomously form a network[6][7][8]. This WSN technology has been applied in various industries including distribution, logistics, construction, transportation, military and health care to realize advance in productivity, safety and human quality of life[9].

In the system being proposed, sensor nodes are attached on livestock to form WSN network and livestock activities are measured for collection and monitoring of biometric data, which is compared with the data of activity changes by livestock diseases stored in a database. If it is over or under the reference values, it is informed to the producer for prompt actions.

Based on the livestock activity data from the system being proposed, it is possible to diagnose health and disease conditions of livestock on time and to be alerted, which will prevent outbreak of livestock diseases in advance and as a result will minimize national economic loss as well as economic and mental damage of livestock farmers.

This paper is configured as follows. Chapter 2 describes the system structure for monitoring livestock activity using WSN, and the service processes provided. Chapter 3 describes measuring method of livestock activity of the system being proposed and evaluates the system performance with a testbed. Lastly chapter 4 wraps up this paper with conclusion.

2 Design of the Proposed Livestock Activity Monitoring System

2.1 System Structure

As illustrated in Figure 1, the proposed system for monitoring livestock activity is configured in three layers of the physical layer, the middle layer and the application layer. The physical layer is composed of environmental sensors collecting environmental data of barns, acceleration sensors measuring livestock activity, GPS modules, CCTVs collecting video images of barns and livestock and environmental control facilities building optimal environment for raising livestock.

The middle layer is composed of the sensor data manager managing collected data from environmental sensors, acceleration sensors and GPS modules of the physical layer, the video data manager managing collected video data from CCTVs, the barn facility manager managing barn facilities, the database storing barn data and livestock activity data, and the management server monitoring barns and livestock.

The sensor data manager formats collected data from environmental sensors, acceleration sensors and GPS modules in the form storable to the database, applies unit conversion appropriate to measured elements and stores the processed data to the database by using update queries.

The barn facility manager operates or manages the environment control facilities by receiving control signals from the livestock and barn management server, and stores such status of barn facilities, the operation time and the number of controls to the database.

The video data manager provides web streams by transmitting videos taken from CCTVs to the livestock and barn management server, classifies them with the barn ID and the camera number and stores them to the database.

The database takes in charge of storing collected data from environmental sensors, acceleration sensors and GPS modules, collected video data from CCTV's, the status of environment control facilities, the operation time and the number of controls, and environmental reference values for automatic control and status alert to corresponding tables.

The management server is placed in between the user and the database, informs data stored in the database to the user periodically, compares the control table of environment control facilities with the environmental reference values stored in the status notification table to automatically control the corresponding environment control facility, or compares and analyzes the existing data of livestock activities store in the database with the measured data of livestock activities to provide the alarm service to the producer in real time via web and SMS if the measured data is over or under the data of activity changes by livestock diseases.

The application layer is composed of application services supporting various platforms including web, PDA and smart phone which are able to provide the livestock biometric monitoring service, the barn monitoring service and barn environment control server to the user.

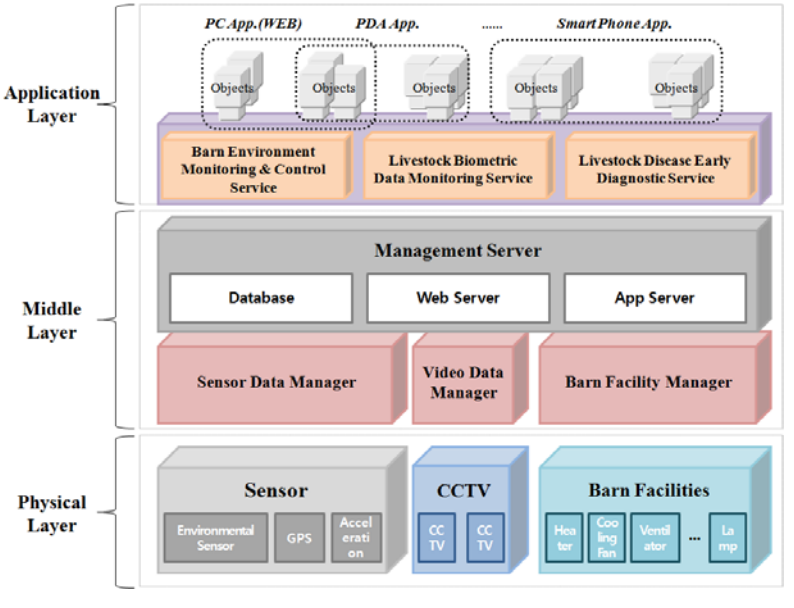


Fig. 1. Livestock Activity Monitoring System Structure

2.2 Service Process

The representative service provided by the system being proposed is the livestock biometric data monitoring and disease early diagnostic service, which is provided by the operation steps as shown in Figure 2.

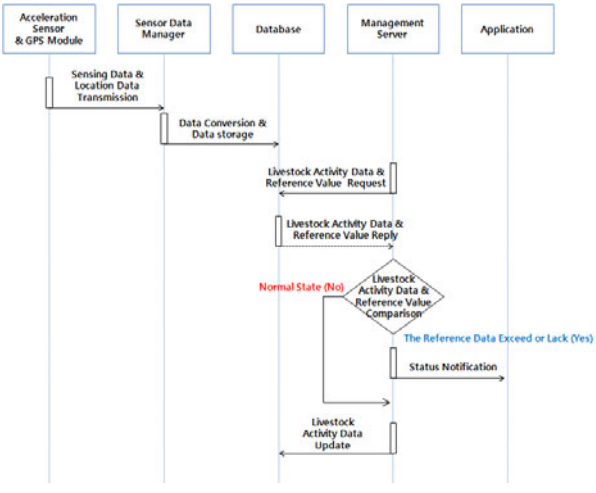


Fig. 2. Service Process of Livestock Biometric Data Monitoring and Disease Early Diagnostic

The livestock activity and location data are measured by the acceleration sensor and the GPS module attached on livestock, then are transmitted to the sensor data manager. The sensor data manager formats collected data in the form storable to the database, applies unit conversion appropriate to the measured element and stores the processed data to the database by using update query.

At this time, the management server calculates livestock activity change by comparing the previous livestock activity data stored in the database and the measured activity data, and then compares it with the data of activity changes by livestock diseases. If normal, the management server updates the database with the collected livestock activity data, and if over or under the reference values, the barn management server calls the producer to alert the status. Through these steps, it is able to identify the livestock health and growth status and increase or decrease in livestock activity caused by diseases in advance.

3 Implementation of the Proposed Livestock Activity Monitoring System

3.1 Livestock Activity Measurement

The methods raising livestock may be divided into indoor raising and outdoor grazing. The system being proposed in this paper measures livestock activity by

using sensor nodes, which acceleration sensors and GPS modules are attached, applicable to both mentioned methods.

When raising livestock indoor, the livestock activity is measured by using acceleration sensors for it is not possible to collect location data from the GPS modules, and when grazing outdoor, the data collected form the acceleration sensors and the location data collected from the GPS modules attached on the sensor nodes are utilized to measure moving distances which give livestock activity data.

Figure 3 shows the sensor node and the GPS module used to measure livestock activity in the system being proposed.

Fig. 3. Sensor Node and GPS Module

3.2 Result

To evaluate performance of the system being proposed in this paper, an acceleration sensor and a GPS module are attached on a mobile robot as illustrated in Figure 4, and a test was performed to measure activity of the virtual livestock.

Fig. 4. Mobile Robot

For the test purpose, the space was limited as shown in Figure 5, sensor nodes were distributed to autonomously form a network, and then physical data obtained from the sensor node mounted on the mobile robot was made to be transmitted to the server system.

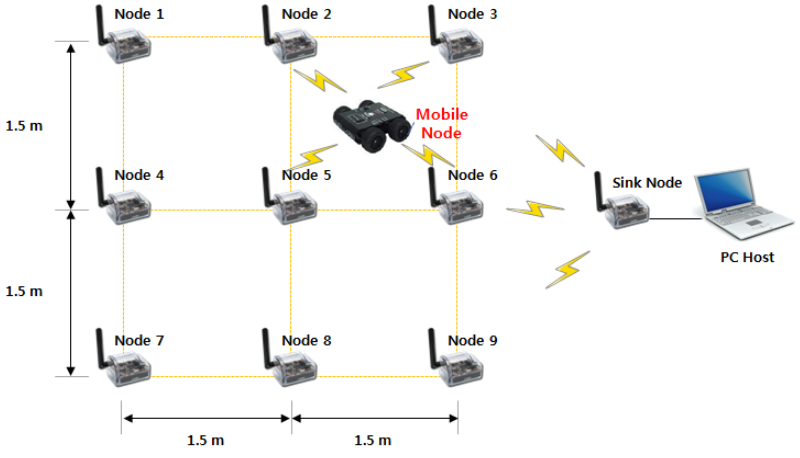


Fig. 5. Wireless Sensor Networks Topology

In addition, in order to provide the livestock activity data collected from the sensor nodes, GUI of the livestock activity monitoring system was developed as shown in Figure 6, through which it was implemented to provide livestock activity data including environmental data such as illuminance, temperature, humidity and CO2, acceleration and position collected from the sensor nodes in the textual and graphical forms.

Table 1. Development Environment

	Type	**Details**
PC Development Environment	OS	Microsoft Windows XP
	Programming Languages	JAVA(JDK 6), C#
Server PC Environment	CPU	Intel Xeon 3.2 Ghz
	RAM	1 GB
	OS	Fedora Linux 2.6.27
	DATABASE	MySQL 5.0
	WAS	Tomcatt 6.0.20
Sensor Node Environment	OS	TinyOS 1.0
	Linux Environment	Cygwin
	Sensor Programming Languages	NesC
	JAVA	JDK 1.4.1

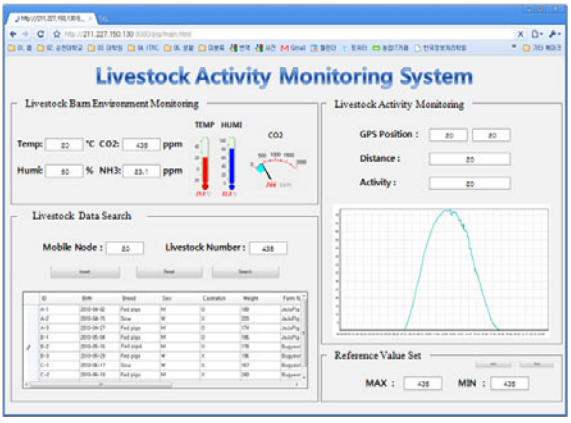


Fig. 6. Livestock Activity Monitoring System GUI

4 Conclusions

Recently in Korea, for the damage caused by the outbreak of livestock diseases such as foot-and-mouth disease and AI has been enormous, and currently in Korea, there is no system for surveillance and early detection of livestock diseases to solve damage caused by them.

In order to minimize such damage, it is necessary to develop collection and analysis technology of livestock biometric data which enables early diagnosis of livestock diseases.

Therefore in this paper, by proposing a wireless sensor network technology applied system which is able to collect and monitor livestock activity, it aims to block the outbreak of livestock diseases in advance, so to minimize national economic loss as well as economic and mental damage of livestock farmers.

The system being proposed attaches sensor nodes on livestock to measure livestock activity, and the collected activity data is compared with the reference data of livestock activity store in the database. If the comparison shows any decrease in livestock activity which may be caused by livestock diseases, the status is notified to the producer in real time, which enables prompt actions to be taken.

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-2011-(C1090-1121-0009)).

This research was financially supported by the Ministry of Education, Science Technology (MEST) and National Research Foundation of Korea(NRF) through the Human Resource Training Project for Regional Innovation.

References

1. Lee, J.-H.: What is a threat to South Korea's livestock industry (Livestock industry research series 11). Focus attention GSnJ No. 55 (2008)
2. 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)
3. Hwang, J.H., Yoe, H.: Study of the Ubiquitous Hog Farm System Using Wireless Sensor Networks for Environmental Monitoring and Facilities Control. Sensors 10, 10752–10777 (2010)
4. Seoul Newspaper, http://www.seoul.co.kr/news/newsView.php?id=20110129002010
5. Kim, J.-S.: Environmental problem and Citizens Science owing to the failure of foot and mouth disease (FMD) policy. The Korean Association for Environmental Sociology 15(1), 85–119 (2011)
6. Akyildiz, I.F., et al.: A survey on Sensor Networks. IEEE Communications Magazine 40(8) (2002)
7. Chong, C.-Y., Kumar, S.P., Hamilton, B.A.: Sensor networks: evolution, opportunities, and challenges. In: Proc. IEEE, vol. 91(8), pp. 1247–1256 (2003)
8. 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)
9. Pyo, C.-S., Chea, J.-S.: Next-generation RFID/USN technology development prospects. Korea Inf. Commun. Soc. 24, 7–13 (2007)

Design of Integrated Control System for Preventing the Spread of Livestock Diseases

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. This study is to develop an integrated control system for preventing the spread of livestock diseases including the food-and-mouth disease (FMD) that has recently been prevalent in South Korea. The system is expected to collect the information and images of livestock and raising facilities by use of radiofrequency identification/ubiquitous sensor network (RFID/USN) and an image processor. Additionally given the fact that livestock diseases may be prevalent, the system is aimed at tracking vehicles carrying feedstuff, milk and excretions by use of radiofrequency identification/global positioning system (RFID/GPS). The collected data are stored in an integrated database (IDB). In particular, the system enables prompt actions against abnormalities that occur in livestock or raising facilities.

Keywords: Livestock, diseases.

1 Introduction

Livestock diseases have been prevalent throughout the world. In the case of malignant diseases, they have been transmitted more rapidly and more damaged to livestock [1]. In particular, FMD is an acute viral infectious disease that infects artiodactyla such as cattle, sheep, goats and hogs [2]. Once an animal is infected, blisters form around mouth, nose and hooves. In some cases, teats may be blistered. The foot-and-mouth disease spreads through various channels such as livestock, people, vehicles, water, feedstuff, air and wild animals. In case FMD occurs, animals within a certain range are slaughtered to prevent its spread (stamping out). In recent times, the Korean government has estimated the loss caused by FMD at about 3-billion dollar, which is about 7 times more than the damage caused by livestock diseases from 2006 until 2009 (about 450-million dollar). Also, avian influenza (AI) caused untold losses [3]. Moreover, the stamping out causes a rise or fall in the price of livestock products, wherefore farmers' sales become lower and consumer confidence shrinks. In addition, it causes damage to backward industries related to feedstuff, raising materials and animal drugs, forward industries related to slaughter and milk processing, and wholesalers and retailers. In the case of Korea, there have been repeated cases where

* Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 169–173, 2011.

it failed in the initial prevention and thus caused spread. Actually, a huge budget was invested whenever the disease occurred. Hereat, there is an urgent need to develop the integrated control system for preventing the spread of livestock diseases.

2 Design of Integrated Control System for Preventing the Spread of Livestock Diseases

The integrated control system consists of an environmental information gathering system, a vehicle tracking system, spread prevention middleware and an integrated monitoring system.

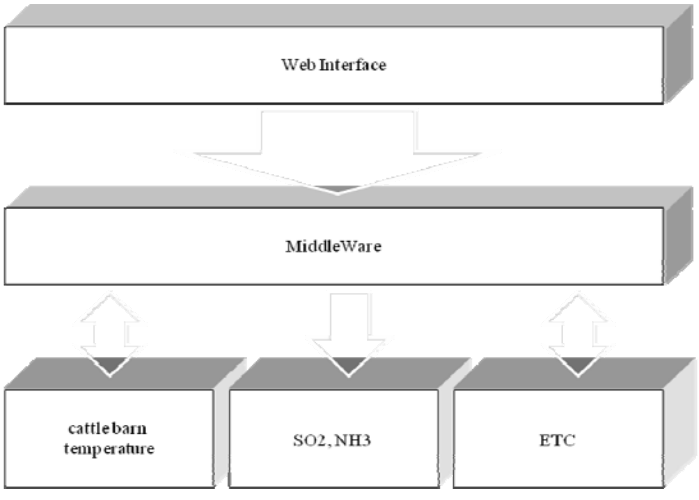


Fig. 1. System Structure

2.1 Environmental Information Gathering System

This system is to gather information on environments, raising conditions, facilities and livestock history from livestock farms and raising facilities all over the country. Moreover, it aims at gathering video information. In this connection, RFID ear tags should be putted on livestock in order to gather livestock information such as gender, weight and feedstuff intake. In addition, the whole process of slaughter should be monitored by livestock farmers, dealers and consumers. Sensors should be placed in raising facilities in order that the temperature and humidity may be checked in real time. Image processors make it possible to remotely take prompt actions against abnormalities that occur in livestock or raising facilities.

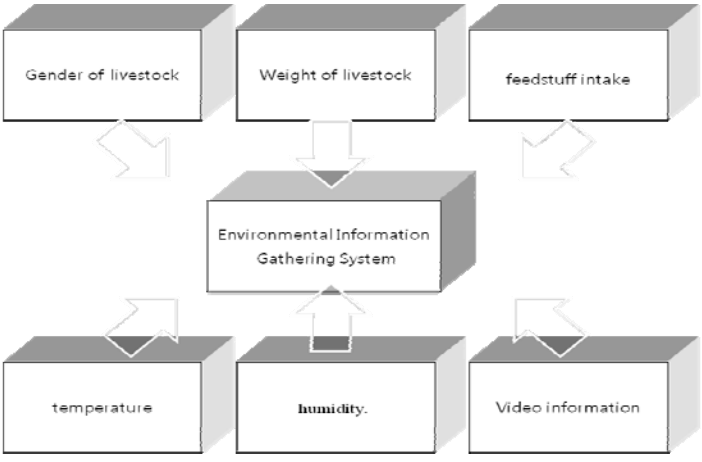


Fig. 2. Environmental Information Gathering System Structure

2.2 Vehicle Tracking System

Livestock diseases are likely to spread along carriageways. Thus, it needs to monitor vehicles that approach raising facilities, e.g., vehicles carrying feedstuff, milk and excretions, in real time, which can be realized by RFID/GPS. 'Fig.3' shows the method to track vehicles. Vehicles, carrying feedstuff, milk and excretions, can be tracked in real time through RFID tags and GPS units.

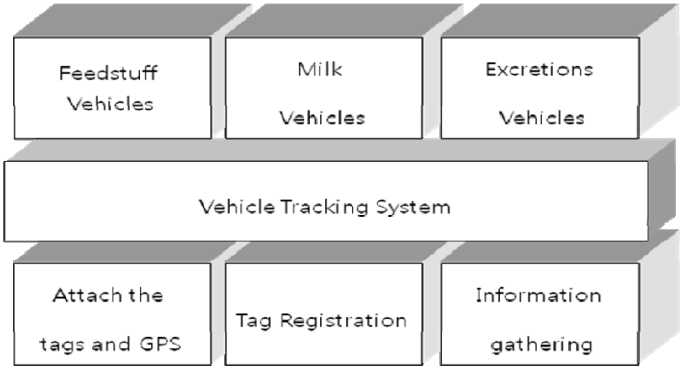


Fig. 3. Vehicle Tracking System Structure

2.3 Spread Prevention Middleware

This middleware consists of a monitoring module, a control module and a database. The monitoring module is to analyze information gathered through the environmental information gathering system (facility information, livestock information, video information, etc.). In case a disease is suspected, it notifies the related veterinarian of the fact and shows the scene in order that he may size the situation up. Additionally, it makes it possible to send a veterinarian to the scene and to issue a livestock disease watch. Add to that, it tracks vehicles that approached the scene of the disease, and restricts the vehicular traffic. The control module is to connect the integrated control center with specific facilities when the integrated monitoring system needs the information of the facilities (video data, facilities, species, etc.). The database is to store information gathered through the environmental information gathering system (facility information, livestock information, video information, etc.), the vehicle tracking system and the monitoring module.

2.4 Integrated Monitoring System

The integrated monitoring system is to monitor information gathered through the spread prevention middleware (temperature, humidity, livestock information, facility information, etc.) and to track the movements of vehicles and livestock, which enables prompt actions at the initiatory stage. Web-based monitoring, user interface and information analysis systems make it possible to provide statistics and information about livestock diseases, preventions and facilities in each region in addition to telemedicine services.

3 Conclusion

This study is to design an integrated control system for preventing livestock diseases. The system tracks the movements of vehicles and livestock in real time, and provides information through the integrated monitoring system, which makes it possible to prevent livestock diseases in advance and to take prompt actions at the initiatory stage. The system is expected to reduce social and economic losses. The study hereafter is to implement the system, and moreover to apply it to raising facilities, and by extension to solve problems.

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-2011-(C1090-1121-0009)).

This research was financially supported by the Ministry of Education, Science Technology (MEST) and National Research Foundation of Korea(NRF) through the Human Resource Training Project for Regional Innovation.

References

1. Hong, S.-G.: A Concept Model for Digitalized Animal Disease Control System. Department of Knowledge Management Graduate School of Information Science Soongsil University (2009)
2. Kim, S.K., Kim, J.E., Paek, D.M.: The Cultural Analysis of 2010-2011 Foot and Mouth Disease Massacre in korea. J. Environ. Health Sci. 37(2), 165–169 (2011)
3. http://biz.heraldm.com/common/Detail.jsp?newsMLId=20110306000143

Hop-Count Based Energy Efficient Traffic Control Mechanism in Wireless Sensor Network*

Yong-Jae Jang[1], Kyoung-Wook Park[2], and Sung-Keun Lee[1,**]

[1] Department of Multimedia Engineering, Sunchon National University, Korea
elsv1114@gmail.com, sklee@sunchon.ac.kr
[2] Division of Culture Contents, Chonnam National University, Korea
kwpark@alex.chonnam.ac.kr

Abstract. In Wireless sensor network, sensor nodes have battery-constrained devices and most of the energy is spent on communication with other nodes. For its traffic feature as burst traffic type toward sink node, it has high probability to network congestion. Network congestion causes packet drop and retransmission of dropped packets draws energy consumption. In particular, the loss of packets that is from sensor node far away from sink node requires additional energy consumption. In this paper, we propose a hop-count based traffic control mechanism that determines packet transfer by considering priority of packet and congestion level as well as hop count. Analysis of proposed mechanism by simulation demonstrated that it improved energy efficiency.

Keywords: Wireless sensor network, Traffic control, Hop count, Queue management mechanism, Energy efficiency.

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. Sensor nodes have limited resource such as limited battery, low calculation capability and so on. Specifically, since it is impossible to change or recharge battery, efficient energy consumption is one of the most important features in WSN[1].

Energy consumption of sensor node is done by environment information sensing, inner data processing, communication between sensors and so on. From these factors, energy consumption for node to node data transfer takes huge portion of entire energy consumption[2]. Therefore, in order to maximize lifetime of sensor network, it requires to reduce unnecessary communication. Since sensor node that runs out of battery can no longer perform the routing and sensing, there are many researches

* 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-2011-(C1090-1121-0009)).

** Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 174–182, 2011.

ongoing to resolve this issue. Another important factor of energy consumption is network congestion.

WSN has low data transfer rate and senses one event in multiple nodes at the same time. Because most traffic is burst to send its data toward sink node, momentary network congestion will occur frequently. As network congestion causes packet-drop and it also causes retransmission of packets, waste of energy is unavoidable. Although protection of network congestion is the best solution for this, it is impossible to avoid congestion in WSN. Hence, it is very important how to deal with congestion efficiently[1][2]. In wireless multimedia sensor networks(WMSNs) that require high transmit rate, it is more prone to happen. So, congestion control is a major focus of research for energy efficiency in WSN[3][4].

In this paper, we propose a hop-count based traffic control mechanism which deals with congestion and guarantees energy efficiency in entire network by considering packet priority and hop count to determine packet retransmission. Through simulation, we analyze the proposed mechanism in comparison with existing mechanisms.

2 Related Works

Wireless Sensor Network is service environment to provide surrounding information or specific information obtained from small sensor nodes in necessary field to user. Sensor nodes are placed where people can't go easily and operated. Hence, They are built wireless sensor network themselves. Sensor nodes are placed in the sensing field, then sensor nodes sense surrounding information and send these information to a sink node. The sink node send received data to user.

WSN's energy, bandwidth, buffer size, processing capability and transmitting current and so on, they are limited. Because people can't reach to sensor node, it is difficult to change or recharge the battery on sensor. And sensor nodes are small and cheap device. They have limited physical resource. Therefore, in WSN, it is difficult to complicated calculate and traffic control process required huge store capability.

Network protocol for efficient traffic control in WSN includes congestion control or reliability guarantee mechanism. Existing protocols can be categorized as three types: congestion control protocol, reliability guarantee protocol and congestion control and reliability guarantee protocol.

CODA[5] is congestion control protocol to detect congestion based on buffer and wireless channel share. GARUDA[6] is reliability guarantee transmission protocol to assure downstream reliability. STCP[7] is one of the WSN transmission protocols, it deals congestion control and reliability guarantee. But these protocols are not considered additional energy for retransmission packets. Thus, they have bad energy efficiency.

3 Hop-Count Based Traffic Control Mechanism

WSN consists of space located sensor nodes as new type of wireless network and a sink node to connect outer network. Data pattern in WSN is a many-to-one type that

multiple sensor nodes deliver data to a single sink node. Each sensor node works not only as a source that generates sensing data but also as a router that broadcasts received data to sink node.

Packets sensed in each node have priority marks. Green mark means high priority, Yellow mark means middle priority and Red mark is low priority. The received node decides received packet process by referring queue state. The queue state is categorized 3 phase: normal operation, congestion avoidance, congestion control. Marked by its importance, packets are processed based on queue occupancy of receiver node. In Normal operation, all the incoming packets are processed. When queue occupancy reaches Red_{min}, the node generates drop probability based on hop count, then drops red and yellow marked packets with the probability. When queue occupancy is up to $\text{Green}_{\text{min}}$, queue state is congestion control state which drops all incoming packets. Fig. 1 illustrates the packet process to management queue.

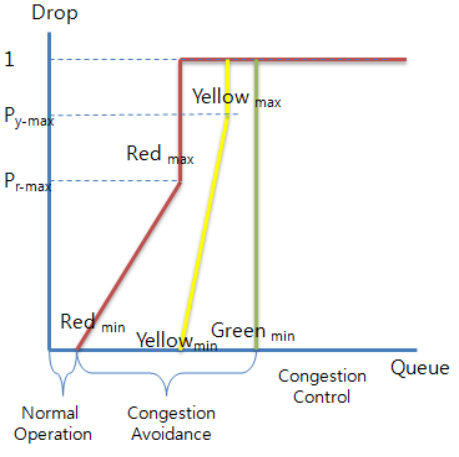


Fig. 1. Packet drop probability

Next formula is to calculate drop probability based on hop count in normal operation and congestion avoidance.

$$P_d = \max(\text{Green}_{\text{min}} - \text{Red}_{\text{max}}, p - D_h/H_{\text{max}}) \; . \tag{1}$$

In formula (1), $\text{Green}_{\text{min}}$ is the minimum queue occupancy that green packets are normally processed. Red_{min} is the maximum queue occupancy that red packets are maximally processed. Standard probability p sets processing standard based on hop of packet. D_h is a constant that presents how far a packet from sender node is delivered to the receiver node. H_{max} is maximum hop count that a node can processed. When packet is dropped in a node far away from the receiver node, hop count is increased; consequently, maximum hop count will be small. Therefore, in the case of receiving a packet that is far away from the receiver node, the node will determine whether it will drop the packet based on ($\text{Green}_{\text{min}}$–$\text{Red}_{\text{max}}$) probability. In the case of receiving a packet that is close to the receiver node, it decides packet drop based on (p–D_h/H_{max}).

```
Algorithm packet_process(packet)
Q_idx : queue occupation
D_h : packet hop count
P_r = rand() mod 10;
if(Q_idx < Red_min)
    Enqueue(packet);
else if(Q_idx ≥ Red_min and Q_idx < Green_min) {
    if(packet.prior == Green)
          Enqueue(packet);
     else if(packet.prior == Yellow or Red) {
          P_d = max(Green_min-Red_max, p-D_h/H_max*10);
         if(P_d < P_r)
             Drop(packet);
          else
             EnQueue(packet);
     }
} else if(Q_idx ≥ Yellow_min and Q_idx < Green_min) {
    if(packet.prior == Green)
          Enqueue(packet);
     else if(packet.prior == Yellow) {
          P_d = max(Green_min-Yellow_max, p-D_h/H_max*10);
         if(P_d < P_r)
             Drop(packet);
          else
             EnQueue(packet);
     }
} else
     Drop(packet);
```

Fig. 2. Queue management algorithm

In the specific hop count, Yellow packet and Red packet have same drop probability. If in the node, process packet's retransmission hop distance is 2 and difference value between $Green_{min}$ and $Yellow_{max}$, $Green_{min}$ and Red_{min} is each 1 and 2, Yellow and Red marked packets have same seed value 2. Therefore Yellow priority is not earlier then Red priority always. But Green marked packets have lowest probability in any circumstance.

Existing protocol mechanism determines packet drop based on queue occupancy. So, when a packet from remote node is dropped, it reduces energy efficiency due to the retransmission that requires additional energy consumption. As it establishes packet drop probability, our mechanism protects from additional energy consumption caused by retransmission and raises higher priority packet transfer probability by applying differentiated packet drop probability based on packet priority, supporting QoS based on application.

4 Performance Analysis and Result

Performance analysis on our proposed mechanism is done by simulation method, and compares existing mechanisms based on queue and RIO[8][9]. Fig. 3 presents the node architecture of simulation. The number of nodes on simulation is total five and the structure of node location is shown as Fig. 3. Node n1, n2, n3 and n5 are to create and broadcast packets and node n4 is only to broadcast packets. Node n4 receives sensing packets of n1, n2 and n3 using congestion control mechanism.

To occur network congestion, it transfers sensing packets of node n5 to n4. All nodes excepted n4 are create 1180 packets. We assume that when packet loss occurs, node n5 automatically retransmits packets to node n4. Source nodes generate sensing data every 500ms. Nodes determine random priority of sensing packets.

Simulation is done to compare and analyze the result in that when packet drop occurs in each node, it stores related packets' hop count and priority information. Then we compare additional energy requirements in terms of packet drop rate and dropped packet retransmission.

In network structure, each node load is n1 = n5 < n2 < n3 < n4. The number of processed packets of node n2, n3, n4 is 1478, 1614 and 2893 respectively. It means n4 is best processing node. Fig. 4 presents additional energy comparison graph for each protocol. The amount of additional energy requirement is calculated using H_{rt} of dropped packet as it assumes energy for 1 hop retransmission on dropped packet in each node to e (energy).

$$E = e \times H_{rt} \tag{2}$$

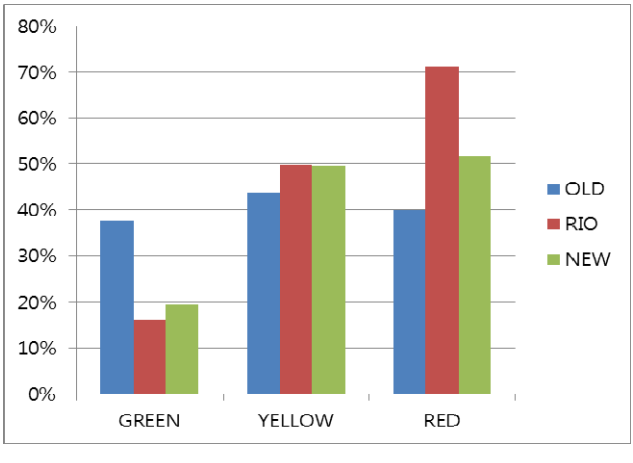
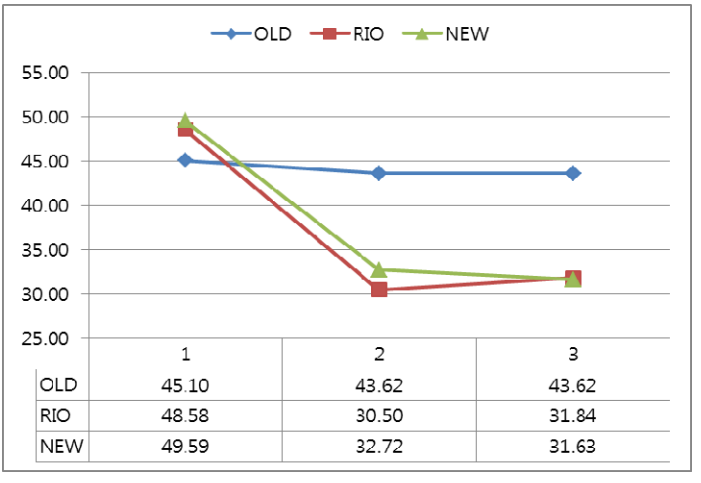


Fig. 3. Average drop probability

The processed packet P_p consist of 1,000 packets S_p created in each node and P_t processed broadcasting packets

$$P_p = S_p + P_t \tag{3}$$

Total processed packets are decided by sensed packets and processed packets in previous node. The additional energy rate is calculated by total processed packets and the amount of additional energy. Fig. 3 presents the average drop rate of each algorithm. In OLD algorithm case, drop rate on each packet priority is Green 37.57%, Yellow 43.73% and Red 39.93%. It means packet drop rate is not considered packet priority. In RIO, Green 16.15%, Yellow 49.74%, Red 71.31%. Proposed algorithm is 19.42%, 49.52%, 51.71% respectively. Drop probability of the Green priority packets is more than RIO, but total drop rate is smaller. Also, proposed algorithm's process rate is more than RIO, and in every nodes, different drop based on packet priority is similar RIO. In node n3, far distant packet drop rate is higher. In this node n3, broadcasting packets are more than sensing packets. It looks case of the broadcasting packets are more than sensing packets. RIO also has same result.



	1	2	3
OLD	45.10	43.62	43.62
RIO	48.58	30.50	31.84
NEW	49.59	32.72	31.63

Fig. 4. Node n4's packet drop probability

Fig. 4 presents graph categorized retransmission packets by retransmission distance in node n4. In OLD algorithm, n4's packet drop rate is each 1 hop 45.1%, 2 hop 43.62%, 3 hop 43.62%. In RIO, 1 hop 48.58%, 2 hop 30.5%, 3 hop 31.84%. Two algorithms dropped packets irrespective of retransmission distance. In n4, proposed algorithm's drop rate is each 1 hop 49.59%, 2 hop 32.72%, 3 hop 31.63%. It means that considered hop count. But drop rates more than RIO. Table 2 presents total energy consumption in node n4. Although RIO's drop rates lower then proposed algorithm, proposed algorithm's total energy efficiency is higher than RIO.

Table 1. Simulation Result

		Old			RIO			New		
diff hop	color	process	drop	drop rate (%)	process	drop	drop rate (%)	process	drop	drop rate (%)
1	Green	489	167	34.15	669	35	5.23	624	66	10.58
	Yellow	498	201	40.36	457	216	47.26	483	235	48.65
	Red	491	192	39.10	434	268	61.75	465	216	46.45
	total	1,478	560	37.89	1,560	519	33.27	1,572	517	32.89

(a) Node n2 packet process

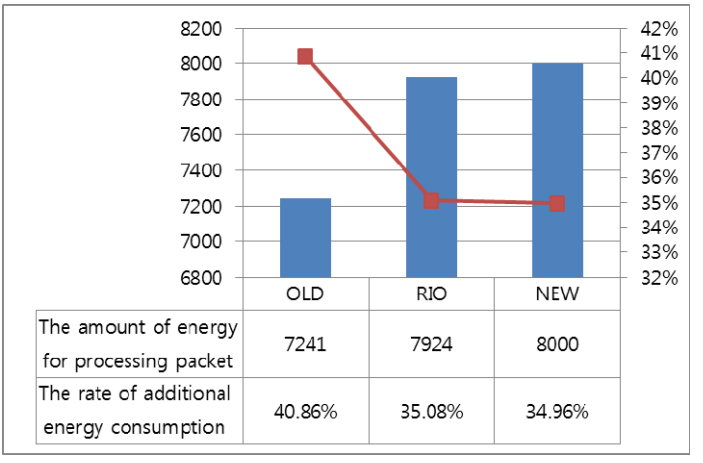
		Old			RIO			New		
diff hop	color	Process	drop	drop rate (%)	process	drop	drop rate (%)	process	drop	drop rate (%)
1	Green	289	96	33.22	556	80	14.39	551	82	14.88
	Yellow	253	102	40.32	211	88	41.71	169	91	53.85
	Red	284	96	33.80	142	102	71.83	178	86	48.31
	total	826	294	35.59	909	270	29.70	898	259	28.84
2	Green	277	102	36.82	574	78	13.59	502	80	15.94
	Yellow	249	103	41.37	203	99	48.77	215	88	40.93
	Red	262	117	44.66	143	104	72.73	214	103	48.13
	total	788	322	40.86	920	281	30.54	931	271	29.11

(b) Node n3 packet process

		Old			RIO			New		
diff hop	color	process	drop	drop rate (%)	process	drop	drop rate (%)	process	drop	drop rate (%)
1	Green	583	256	43.91	677	152	22.45	679	186	27.39
	Yellow	591	275	46.53	537	317	59.03	533	333	62.48
	Red	633	284	44.87	507	367	72.39	516	338	65.50
	total	1,807	815	45.10	1721	836	48.58	1,728	857	49.59
2	Green	118	43	36.44	264	52	19.70	267	65	24.34
	Yellow	97	43	44.33	81	39	48.15	52	26	50.00
	Red	115	42	36.52	32	24	75.00	63	34	53.97
	total	788	322	43.62	377	115	30.50	382	125	32.72
3	Green	110	45	40.91	278	60	21.58	239	56	23.43
	Yellow	97	48	49.48	71	38	53.52	80	33	41.25
	Red	91	37	40.66	31	23	74.19	73	35	47.95
	total	298	130	43.62	380	121	31.84	392	124	31.63

(c) Node n4 packet process

This result means that our proposed algorithm guarantees transmission rate per each priority and reduces additional energy requirement for packet retransmission.



	OLD	RIO	NEW
The amount of energy for processing packet	7241	7924	8000
The rate of additional energy consumption	40.86%	35.08%	34.96%

Fig. 5. Energy sending for retransmission packet

5 Conclusions

It is important issue to save sensor node energy in WSN. Retransmission of dropped packets causes serious energy consumption. Existing algorithms drop packets according to queue occupation and support QoS based on packet priority. It is inefficient for energy consumption.

Our algorithm considers not only packet priority congestion level of network but also transmission distance of packet. The farther the received packet's retransmission distance is, the lower the packet drop rate is resulting enhancement of energy efficiency. Simulation result shows this fact.

Next issue for research is to apply our proposal to a large scale network environment to perform its energy efficiency and QoS guarantee.

Acknowledgement. "This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2011-0014900)".

References

1. Akylidiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: A Survey on Sensor Networks. IEEE Communication Magazine 40(8), 102–114 (2002)
2. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless Sensor Networs: A survey. Computer Networks 38(4), 393–422 (2002)

3. Akylidiz, I.F., Melodai, T., Chowdury, K.R.: A Survey on Wireless Multimedia Sensor Networks. Computer Networks (Elsvier) 51(4), 921–960 (2007)
4. Ehsan, S., Hamdaui, B.: A Survey on Energy-Efficient Routing Techniques with QoS Assurances for Wireless Multimedia Sensor Networks. IEEE Communications Surveys and Tutorials 99, 1–14 (2011)
5. Wan, C.-Y., Eisenman, S.B., Campbell, A.T.: CODA:Congestion Detection and Avoidance in Sensor Networks. In: 2003 ACM Conference on Embedded Networked Sensor System (Sensys 2003), Los Angeles, CA, pp. 266–279 (2003)
6. Park, S.-J., Vedantham, R., Sivakumar, R., Akyildiz, I.F.: A Scalable Approach for Reliable Downstream Data Delivery in Wireless Sensor Networks. In: 2004 ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2004), Roppongi, Japan, pp. 78–89 (2004)
7. Iyer, Y.G., Gandham, S., Venkatesan, S.: STCP: A Generic Transport Layer Protocol for Wireless Sensor Networks. In: 14th IEEE International Conference on Computer Communications and Networks (ICCCN 2005), San Diego, CA, pp. 449–454 (2005)
8. Mellia, M., Toica, I., Zhang, H., Torino, P.: Packet Marking for Web Traffic in Networks with RIO Routers. In: IEEE Global Telecommunications Conference, GLOBECOM 2001, vol. 3, pp. 1828–1833 (2001)
9. Floyd, S., Jacobson, V.: Random Early Detection Gateways for Congestion Avoidance. IEEE/ACM Trans. Networks 1(4), 397–413 (2003)

An Energy-Efficient Routing Algorithm in Wireless Sensor Networks*

Yong-Jae Jang[1], Si-Yeong Bae[2], and Sung-Keun Lee[1,**]

[1] Dept. of Multimedia Engineering, Sunchon National University
elsv1114@gmail.com, sklee@sunchon.ac.kr
[2] Dept. of Computer Science, Sunchon National University
bsy233@sunchon.ac.kr

Abstract. Network architectures and protocols are important aspects in the design of wireless sensor networks(WSNs). Due to the severe energy constraint of sensor nodes, network architectural design has a big impact on the energy consumption and thus the operational lifetime of the whole network. To prolong the lifetime of WSNs, energy efficiency should be considered. We propose an energy efficient routing algorithm that establishes routing table using broadcasting which is cyclically transmitting and conduct performance analysis with simulation method.

Keywords: Wireless Sensor Networks, Routing Protocol, Broadcasting, Energy Efficient.

1 Introduction

Wireless Sensor Networks (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 node [1]. A WSN typically consists of a large number of sensor nodes that are deployed in a region of interest. These sensor nodes have limited processing, storage capacities and limited energy capacity. Sensor nodes are usually powered by battery. In most situations, they are deployed in a harsh or hostile environment, where it is very difficult or even impossible to change or recharge the batteries. To prolong the operational lifetime of a sensor network, energy efficiency should be considered in every aspect of sensor network design[2].

This study proposes the routing protocol that reflects the network state and reduces energy consumption in limited hardware resources of WSNs.

2 Related Work

Traditional routing algorisms for WSNs are introduced as follow.

* This research was supported by the MKE, Korea under ITRC support program supervised by the NIPA (NIPA-2011-(C1090-1121-0009)).

** Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 183–189, 2011.

Low-Energy Adaptive Clustering Hierarchy (LEACH)[3] is based on an aggregation technique that combines or aggregates the original data into a smaller size of data that carry only meaningful information of all individual sensors. For this purpose, LEACH divides a network into several clusters of sensors, which are constructed by using localized coordination and control not only to reduce the amount of data that are transmitted to the sink, but also to make routing and data dissemination more scalable and robust. Given that the energy dissipation of the sensors depends on the distance and the data size to be transmitted, LEACH attempts to transmit data over short distances and reduce the number of transmission and reception operations.

Directed diffusion[4] is technique for disseminating queries across the entire sensor network by flooding. In directed diffusion, the data generated by a sensor is identified by an attribute value pair. The query answer generated in response to a named interest is data centric. Directed diffusion frequently exploits the spatio-temporal correlation in the data streams and is more applicable for dense, static sensor networks. Therefore, they are not highly suitable to the low-correlation situations of other kinds of networks, for example, mobile sensor networks.

Sensor Protocols for Information via Negotiation(SPIN)[5],[6] protocols were designed in a way to improve classic flooding protocols and overcome the problems they may cause, for example, implosion and overlap, which were discussed earlier. In addition, flooding, when used, makes the sensors blindly consume their available resources. The SPIEN protocols are resource aware and resource adaptive. The sensors running the SPIN protocols are able to compute the energy consumption required to compute, send, and receive data over the network. Thus, they can make informed decisions for efficient use of their own resources.

3 Proposal of Protocol

A WSN consists of a large number of sensor nodes, which are densely deployed. Each nodes have state change frequently. The protocol should properly respond to the change, operate in limited resource of sensor nodes and reduce energy consumption.

Most data pattern in WSN is a many-to-one type that sensor nodes deliver data to a single sink node. The data that transmitted to the sink is among the broadcasting, multicasting, geocasting packet that is aimed to query, program patch and control information transmit. These data have differentiated transmission period and interval that depend on application service. Therefore, the sensor nodes can only operate with the path to sink.

The routing of each node refers to its routing table. The routing table is updated in 24 hours period and constructed by using broadcasting information. The message that be broadcasted by the sink increases the hop of message whenever the message pass a node. Each node analyzes ID of the broadcasting message that received from neighbor node. If node received the message of same ID more than 3 times, node determines

the path priority based on hop count and energy level of sender and updates its routing table. Table 1 shows fields of broadcasting message for routing.

Table 1. Fields of broadcasting message for routing

Field	Signification
ID	Broadcasting message Identifier
Flag	To classify broadcasting message into three : first routing construct message, routing table update message and normal broadcasting message
Node Identifier	Sender identifier
Location Information	Sender location information
Hops to Sink	Hop count from source to sink
Energy level	Residuary energy level of sender

```
Initialize:
  brdcst_count = 0;
  tmp_brd;
  MAX = 2;
if(!tmp_brd.empty()){
  tmp_brd.insert(brdcst_packet);
  ++brdcst_count;
  Timer.Start();
  modify(&brdcst_packet);
  broadcast(brdcst_packet);
}
else if(brdcst_count == MAX){
  Timer.Stop();
  r_tableUpdate();
}
else{
  if(!tmp_brd.comp(brdcst_packet)){
    tmp_brd.insert(brdcst_packet);
    ++brdcst_count;
    Timer.Restart();
    modify(&brdcst_packet);
    broadcst(brdcst_packet);
}}
If(Timer.done){
  r_tableUpdate();
}
```

Fig. 1. Algorithm for updating routing table

First, sink is to disseminate broadcast packet to neighbor nodes. Each node that are received the packet compares ID of the received packet with ID of data in temporary memory. Node may receive broadcast packet many times through other paths. Node counts the packets that have same ID and controls energy cost due to duplicate

broadcast packet. The node stores received broadcast packet to temporary memory and increases counter. When ID of message in temporary memory is same the received packet's it, node compares its node identifiers. If node received same ID and node identifier, received packet is redundancy. In this case, node drops received broadcast packet. If counter in node reach predefined value MAX, node determines the path priority based on temporary memory, updates routing table and drops all broadcast packets that should receive from neighbor nodes. Some nodes are waiting for the packet to receive because the value of counter in the node does not reach MAX. To avoid this situation, nodes set 'receive waiting timer' when nodes have received broadcast packet. If the value of counter in the node does not reach MAX, node updates routing table based on temporary memory.

Fig 1 is algorithm for routing table construct. brdcst_count is broadcast packet counter. tmp_brd is class object that managed received broadcast packet. The class include the following several methods: empty() checks temporary memory whether temporary memory is empty. insert() stores received broadcast packet in temporary memory and determines path priority based on hop count and energy level. comp() compares packet of temporary memory with received broadcast packet. MAX is the predefined value for received broadcast packet counter. If receive counter is max, node drops all packets that should receive from neighbor nodes. modify() alters node identifier and hops to sink in received broadcast packet to broadcast neighbor nodes. Fig 2,3 depict routing table construct mechanism.

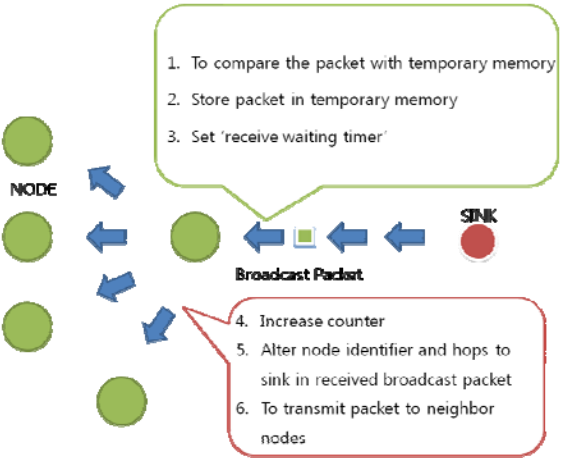


Fig. 2. Routing table construct mechanism 1

In Fig 3, Node 2 receives packets in same order as Node 1, Node 3. Node 2 has received broadcast packet 3 times and updates routing table. However, Node1 and Node 3 can't receive broadcast packet 3times. Each node has 'receive waiting timer'. If the node can't receive broadcast packet 3times when 'receive waiting timer' is done, node updates routing table based on temporary memory.

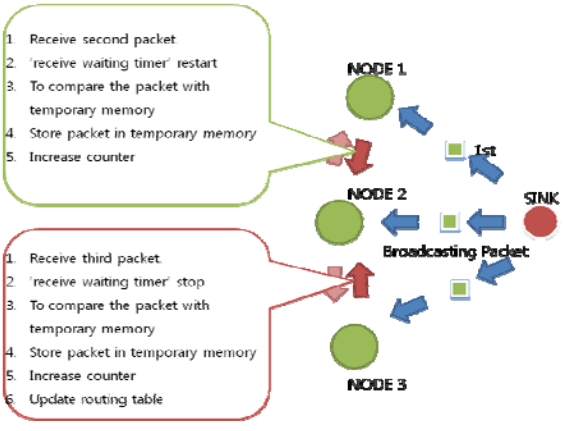


Fig. 3. Routing table construct mechanism 2

4 Performance Analysis and Result

4.1 Experiment Configuration

Performance analysis for our proposed mechanism has conducted using simulation method. Fig 4 depicts node structure design in our simulation. Simulation was carried on under Intel Core 2 Quad Q9550 2.83GHz, a CPU, RAM 8GM, Windows 7, an operating system, Visual studio 2010.

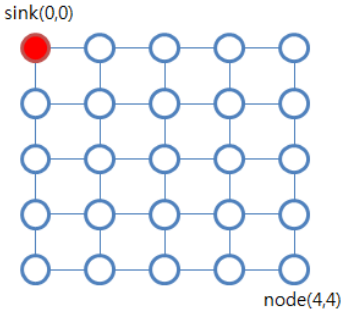


Fig. 4. Configuration of nodes

Simulation has 25 nodes that are formed 5x5 grid construction. Node n(x,y) has only four direct dissemination: n(x±1,y) and n(x,y±1). Sink transmits the broadcast packet for the routing table update. The value of 'receive waiting timer' is set to 3 seconds. In simulation, nodes only send received packet. The factor determining path priority is limited as hop count. To analyze the result, the result concludes node

ID(location information), congestion level, next node and hops to sink of each nodes. If routing table update of all nodes is done, we analyze routing tables of all nodes.

4.2 Simulation Result and Analysis

Fig 5 shows routing table of edge nodes that is n(0,4), n(4,0), n(4,4). Edge node updates routing table when 'receive waiting timer' is done because edge node has only 2 neighbor nodes. Node n(4,4) has paths that have same hops to sink.

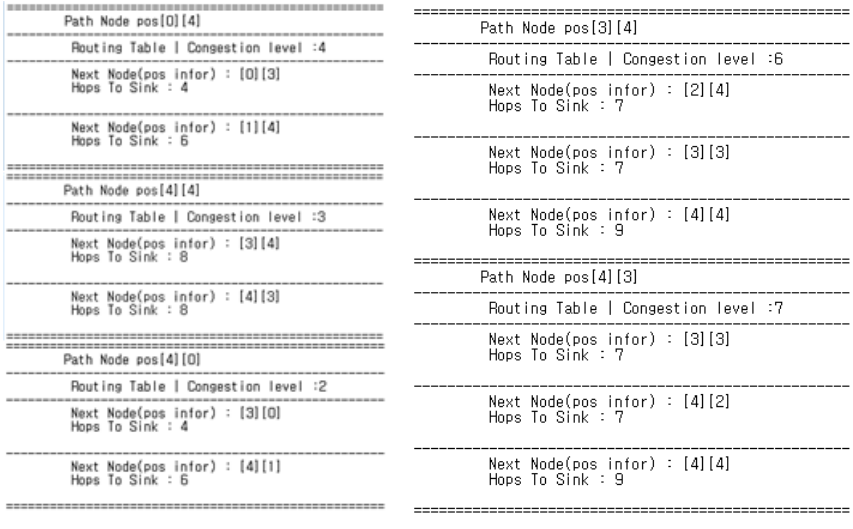


Fig. 5. Routing table of edge nodes and nodes that has same hops to sink

Node n(3,4) and n(4,3) have same hops to sink but congestion level of n(3,4) is lower than n(4,4). Therefore, our proposed mechanism reflects not only hop count but also congestion level in portion of network. Thus, our proposed mechanism can support energy efficient and proper routing service for network state by constructing the routing table based on hop count and congestion avoid.

5 Conclusions

WSNs suffer from the limitations of several network resources, energy is the most crucial resource because it determines the lifetime of a sensor. Therefore, algorithms designed for sensors should be as energy efficient as possible to extend their lifetime. This paper proposes an energy efficient routing algorithm by using broadcasting. Proposed algorithm constructs optimal path based on energy level and hop count and can perform dynamic routing by managing the number of times being received. This paper proposes network protocol algorithm that reduces excessive resource cost by counting broadcast packet.

Acknowledgement. "This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2011-0014900)".

References

1. Akylidiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: A survey on Sensor Networks. IEEE Communication Magazine (2002)
2. Akkaya, K., Younes, M.: A survey on routing protocols for wireless sensor networks. Ad Hoc Networks 3(3), 325–349 (2005)
3. Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: An application-specific protocol architecture for wireless microsensor networks. IEEE Transactions on Wireless Communications 1(4), 660–670 (2002)
4. Intanagonwiwat, C., Govindan, R., Estrin, D., Heidemann, J.: Directed diffusion for sensor networking. IEEE/ACM Transactions on Networking (TON) 11(1), 2–16 (2003)
5. Heinzelman, W.R., Kulik, J., Balakrishnan, H.: Adaptive protocols for information dissemination in wireless sensor networks. In: Proceedings ACM MobiCom 1999, Seattle, WA, August 1999, pp. 174–185 (1999)
6. Kulik, J., Heinzelman, W., Balakrishnan, H.: Negotiation-based protocols for disseminating information in wireless sensor networks. Wireless Networks 8(2/3), 169–185 (2002)

A Study on SOAP-Based Standard Platform for the Connection Activation of Report Management Systems

Hongro Lee, Yongju Shin, Jeongkyum Kim, Ki-Seok Choi, and Jae-Soo Kim

NTIS div., Korea Institute of Science & Technology Information(KISTI)
{hongro,yjshin,mcjg82,choi,jaesoo}@kisti.re.kr

Abstract. In May 2008, the Korean Government has earmarked KISTI (Korea Institute of Science & Technology Information) as the R&D report management and distribution exclusive institute. Therefore, the KISTI have been building a database through a collection of R&D reports from the number of 125 the research management specialized organization. In this paper, we designed and implemented a standard platform based on SOAP for the revitalization of reports for linking between the research management specialized organization and the report R&D output management and the distribution exclusive institute. The proposed standard platform is expected to contribute as providing a service to customers in the research management specialized organization for effective registration and management of the national R&D reports.

Keywords: National R&D Reports, government, outcome, database, SOAP, RESTful, web service, standard platform.

1 Introduction

The Korean Government have been fulfilled a study on systematic management and the promotion of effective utilization for the research outcome generated by the national R&D business since September 2007. The KISTI be designated as the R&D output management and distribution exclusive institute by the National Science & Technology Commission in May 2008. However, in spite of these efforts of the Korean Government, each research management specialized organization still has different schemas and data types of outcome management systems. Also, the human resource and time is being wasted because their retained report is huge. Therefore, the KISTI, the R&D output management and distribution exclusive institute, is trying to increase a utilization of standardized reports to the research management specialized organization. In this paper, we designed and implemented a standard platform based on SOAP for the revitalization of reports linking between the research management specialized organization and the report R&D output management and distribution exclusive institute. Firstly, we review the previous works and applications on traditional web service field. There are two trends that are the SOAP based technology, the RESTful based technology.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 190–195, 2011.

1.1 The SOAP-Based Technology

Generally, the Web Service technology can be defined a standard technology for sharing an information which opens to abstract Service type from distributed digital contents. Also, the SOA (Service Oriented Architecture) is a representative technology. SOAP (Simple Object Access Protocol) based Web Service technology began in demands of corporation which freely want to use a service in remote method such as the API by messaging standard information with SOAP to link between applications on each other platforms [2]. All messages generated by the SOAP represent a XML document consist of SOAP Envelope, SOAP Header and SOAP body in Figure 1. Therefore, the SOAP demands an environment for development because of high development level of difficulty such as complex encoding and decoding process.

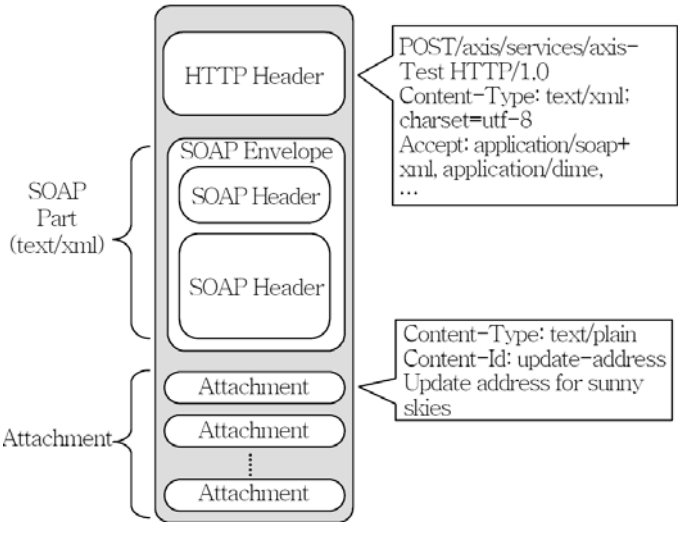


Fig. 1. A Structure of the SOAP web service

1.2 The RESTful-Based Technology

The REST is a network based architecture proposed by Roy Fielding who is an originator of Web. It can transfer the data by HTTP protocol without the additional session management or the network layer. Also, it has an advantage when developer implement an application simple and performance and scalability can be increase by strict separation of components parts between client and server. However, the RESTful Service is independent in standardized infrastructure for development of applications compared with the SOAP based Web Service well-equipped infrastructure. Therefore, the RESTful based technology has a problem in developing the standardized application and the maintenance of developed system because of the absence of standard for the REST.

1.3 Comparison of the Web Services for Standard Platform Design

The SOAP-based Web Service is a technology for improving the service performance using WSDL and RESTful Web Service is a technology for sharing the resource based on URI. The RESTful Web Service directly provides a resource to end user without the middleware which register and store resource, while the SOAP based Web Service depend on SOA infrastructure.

2 Design of Standard Platform for National R&D Reports

In order to develop the standard platform for registering and distributing of the national R&D reports, we should firstly consider a selection of the standardized and practical architecture in each Research Management Specialized Organization. As we mentioned in chapter 2.2, the SOAP based Web Service has an advantage in a standardization and independence from platform compared with the RESTful based Web Service. Although SOAP based Web Service has a difficulty due to complex structure at the development of application, if a standard-platform provider timely update a resource, the Research Management Specialized Organization can use service without regard to complexity of architecture.

Secondly, we should design our system consider with an information security because the standard platform have to exchange important data between public institutions. However, the SOAP and RESTful based service unfortunately does not include their security layer. Therefore, in case of SOAP, the packet was transmitted on the HTTPS Protocol and they need additional SSL (Secure Socket Layer) or WS layer. Also, we have to implement the authentication or the IP blocking function for strengthening security of SOAP server. Finally, REST based service lacks security because of dependent on HTTP.

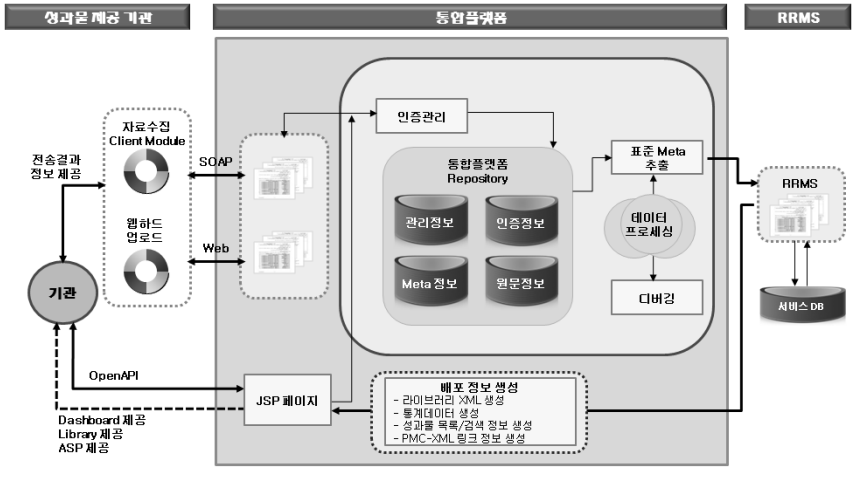


Fig. 2. Block diagram of proposed method

Therefore, we propose a design of the standard platform for the national R&D report using the SOAP based Web Service. Additionally, we apply the authentication and the IP blocking function for strengthening an information security. The structure diagram of the proposed method is shown in Fig. 2.

3 Experimental Results

The proposed standard Platform implemented to two modules included a registration module can collect the national R&D report from the Research Management Specialized Organization with efficiency and a distribution module can easily download pre-registered data. The registration module has a transmission function of meta-data and files using the Flex. Also, the protocol of data transmission was applied with the SOAP Web Protocol for transmitting the data against the firewall. Data transmission method be used the JAX-WS of AXIS2, iBaits 3.0 as the data storage method. The main function of our system has a subject mapping function for the multi-file, a temporary storing function against the transmission error, a directory listing function, a registration function for the batch files. Our registration module is shown in figure 3.

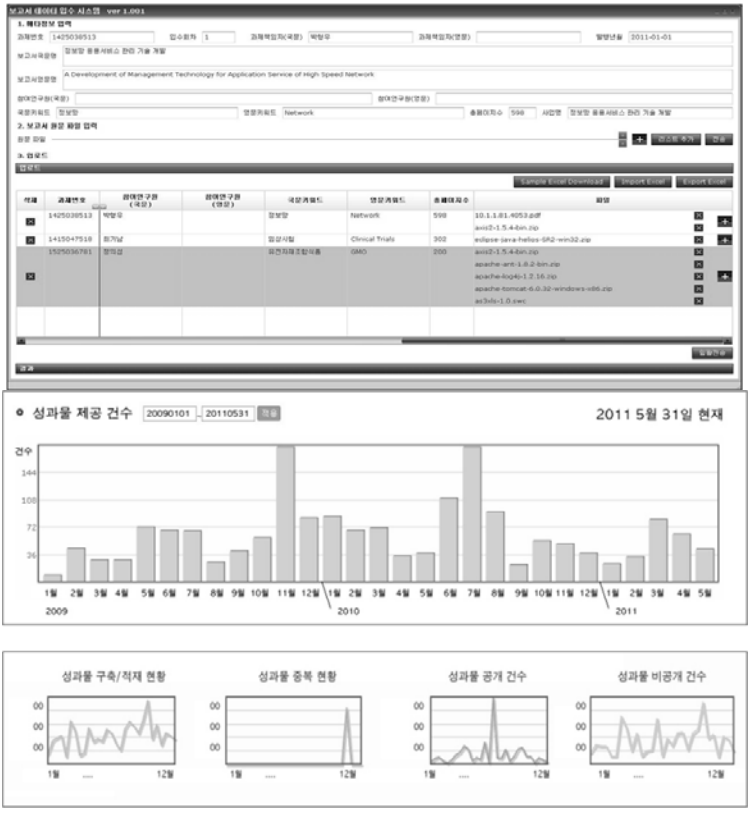


Fig. 3. Implemented Distribution Module

The distribution module can be divided into the individualized Service, the distribution Service, ASP service. The individualized service consists of the viewing function that user in the registration organization can check on a building process of database and a statistics function for analyzing the pre-registered report with periods. Also, user in public organization can directly download the constructed report file using the distribution service. Finally, the ASP Service is a data linking service of open type based on Open-API applying with linking for registered report information by external user. The distribution module is shown in figure 3.

Finally, we implemented an application applied with our slide retrieval method to test the performance as shown in Figure 2. Since flexible formation and modification of queries can only be obtained by involving the user in the retrieval procedure for content-based image retrieval, we designed crucially the application. User interfaces in our systems typically consisted of a query formulation part and a result presentation part.

4 Conclusion

In this paper, we designed and implemented a standard platform based on SOAP for the revitalization of reports linking between the research management specialized organization and the report R&D output management and distribution exclusive institute. The proposed standard platform is expected to contribute as providing a service to customer in the research management specialized organization for an effective registration and management of the national R&D reports. In the future, the Eight R&D output management and distribution exclusive institute will reflect requirements from research management specialized organization.

References

1. Lee, H.-R., Kim, J.-K., Choi, K.-N., Choi, K.-S., Kim, J.-S.: A Study on the Improvement of Framework for Collection and Management of National R&D Reports. Korea Society for Internet Information, pp. 297–302 (June 2010)
2. Park, Y.: Comparison of SOAP based Service and RESTful Web service Technology. ETRI 25(2), 112–120 (2010)
3. Fielding, R.: Architectural Styles and the Design of Network-based Software Architectures. Dissertation of Doctor of Philosophy in Information and Computer Science, University of California, IRVINE (2000)
4. Kim, C.: Web Service Concept and Application Service Trends. National IT Industry Promotion Agency, Weekly Trends, no. 1395.9, pp. 13–26 (May 2009)
5. Richardson, L., Ruby, S.: RESTful Web Services. O'Reilly Media (May 2007)
6. Pautasso, C.: REST vs. SOAP: Making the Right Architectural Decision. In: WWW 2008, April 2008, pp. 805–814 (2008)
7. Hines, G.: RESTful Web Services and Drupal. PingVision (2008)

8. Fensel, D., Kerrigan, M.: Implementing Semantic Web Services. Springer (2008)
9. Verma, K., et al.: METEOR-S WSDI: A Scalable Infrastructure of Registries for Semantic Publication and Discovery of Web Services. J. of Information Technol. and Management, Special Issue on Universal Global Integration 6(1), 17–39 (2005)
10. Hepp, M.: Semantic Web Based E-Commerce: The Good Relations Ontology. In: Semantic Web Technology Conference (June 2009)
11. Tim, B.-L.: Putting Government Data online (June 2009)

The 4-Tier Design Pattern for the Development of an Android Application

Woon-Yong Kim and Seok-Gyu Park

Dept. of Computer and Internet Technique, Gangwon Provincial College, Korea
{wykim,skpark}@gw.ac.kr

Abstract. Recently, increasing in demand for smart phone and related applications has changed people's life style rapidly. In particular, Apple's IPhone and Google's Android phone has been increased in demand, and the application development of those has become the necessary elements in the business area. But because the application of the smart phone has made in the special framework environment and the design of the application has made base on the user interface with domain area, the application complexity increases. For reduce the complexity, we propose the 4-Tire design pattern for the android application. In this design pattern, we separate the area of the executable components (Activities, Services, Broadcast Receiver, and Content Providers) and domain area for the low coupling. This structure can improve the ability of the maintenance and reduce the cost of the application development.

Keywords: Mobile, Application, Design, Pattern, Smart Phone, Android.

1 Introduction

With the rapid growth of mobile environment, increasing the demand of smart phones and related applications has been changing the paradigm of life style. Especially, Apple's IPhone and Google's Android phone has been increased in demand, and the application development of those has become the necessary elements in the business area. But because the application of the smart phone has made in the special framework environment and the design of the application has made base on the user interface with domain area, the application complexity increases so that this model make it difficult for the upgrade and inhibit analysis capability. In this paper, we propose a design pattern for android for increasing the efficiency of the application development. This pattern has the 4-Tire design layers for the android application. We separate the area of the executable components (Activities, Services, Broadcast Receiver and Content Providers) and domain area for the low coupling. And then, we connect with interface class between the executable components and domain area and use the Factory Design Pattern for the flow among contents. And also, we compose domain area with the unique features. This approach can make loose coupled structure for the application. This structure can improve the ability of the maintenance and reduce the cost of the application development.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 196–203, 2011.

The rest of this paper is organized as follows. In section 2, we briefly review related work for the application design and the environment of the application for the android system. In section 3, we propose the 4-Tier design pattern for the android system. In section 4, we show typical applications with the proposed design pattern. Finally, we conclude in Section 5.

2 Related Work

2.1 Android Application Structure

Android application can make use of elements of other applications provided those applications permit it. For this to work, the system must be able to start an application process when any part of it is needed unlike applications on most other systems. And also Android applications don't have a single entry point for everything in the application and have essential components that the system can instantiate and run as needed [1][2]. The application consists of Manifest file that descript everything in the application, application components and resources. Fig. 1 shows elements of the application.

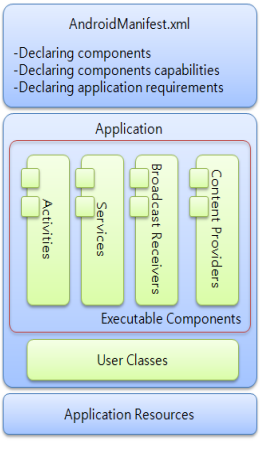


Fig. 1. The elements of Android application

There are four types of executable components in the android application such as Activities, Services, Broadcast Receivers and Content Providers. Those components have the life cycle of their own and executable model. And also they have connection structure with components and they act alone or with other components freely. Generally, application execution process is made by the call of the executable components so that the process of design for the android application depends on the action of the executable components. Fig.2 shows the connection among tasks, process, package and android components.

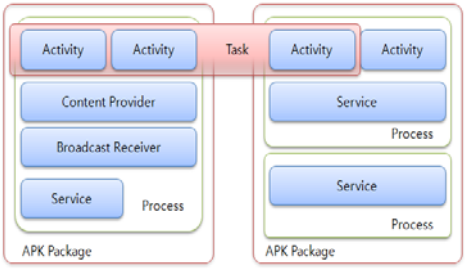


Fig. 2. The connection among task and process and components

2.2 The Design for the Mobile Application

There are various researches of software design for mobile platforms. Those researches have used Agile method for mobile environments generally. XP(eXtreme Programming) is a typical methodology. This approach execute design and implementation based on four principles such as simplicity, interactions, communication feedback and courage [3][4]. And also, MASAM(Mobile Application SW Based Agile Methodology) that is an extended software development methodology for mobile platform include the concepts of an architecture based on model and use the Agile methodology and patterns. and the methodology proposed the steps of preparation, refinement, development and commercialization[5]. And Mobile-D that is a methodology to complement the technical constraints of the mobile environment is based on eXtreme Programming, Crystal Methodologies and RUP(Rational Unified Process). This approach proposed each steps and practices [6]. In Natchetoi's Service-oriented architecture for mobile applications, they proposed design technique of SOA-based architecture for mobile devices by considering their characteristics [7].

The proposed mobile-based application development process offered general mobile application development process based on Agile methodologies and show the abstract approach with the overall process. But the application with special framework has itself feature and restrictions and act with their life cycle so that we need to reflect those features. And also, the android application made by components based on user interface design. This approach can be an independent domain model. In this paper, we propose the detailed design methodology named 4-tire design pattern for the android application that has the popularity in the smart phone area. The application of this approach in the design process will create a low coupling among the objects and will make maintenance easier.

3 The 4-Tier Design Pattern for the Android Application

Android application design can be consisted of executable components that have their life cycle and the functions for the user interface. In this environment, the design included user domain increases the complexity of the design and make it difficult to

create flexible code. In this paper, we divide into four levels and give the unique features of each layer through considering the unit-feature. It can be reduce the complexity and preserve the low coupling.

3.1 The Proposed 4-Tier Structure

The proposed 4-tier structure is shown in Fig. 3. It consists of component layer, interface layer, implement layer and persistency layer.

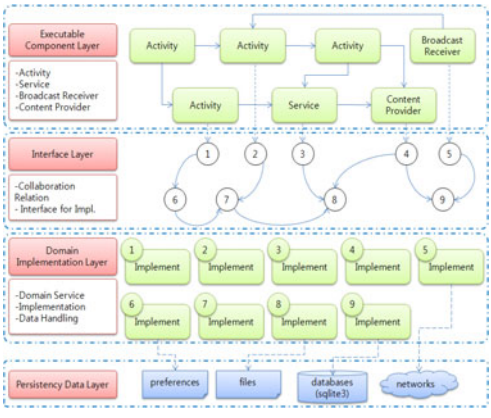


Fig. 3. The 4-Tier structure for the android application

Executable Component Layer

The components in the android system divide into Activity, Service, Broadcast Receiver and Content Provider. Activity represent user interface that can be faceless, can be in a floating window and can return a value and can be embedded. Service component is faceless classes that run in the background such as network download and run in your application's process or their own process. Broadcast Receiver component does nothing but receive and react to broadcast announcements that the battery is low and a picture has been taken so on. Content Provider component enables sharing of data across applications such as address book, photo gallery so on. And also, the component uniform API for querying, delete, update, insert rows. Those components have their own life cycle and can call each other components. So we need to separate components and domain area because including user domain in the component makes program structure complexity in this component environment. It can be construct a definite structure by presenting the relations of components and unique features without domain area.

Interface Layer

Interface Layer provides connection between executable components and domain area and keeps a low coupling. This Interface layer includes implements class method information and present relations among the interface methods. It can recognize the flow and operation of the application easily.

Domain Layer
Domain layer that is user domain area has implement classes. The implement classes associate the interface layer and compose classes independently and have only the domain features. It can make the classes flexible to update and modify the user requests.

Persistency Data Layer
Persistency data layer has an environment information of applications, file system, database and network environment. Implement classes use this layer to read or save data.

3.2 The Relationship with Each Layer

Each component and classes in the layers have association closely. This relationship is shown by class diagram in the Fig. 4.

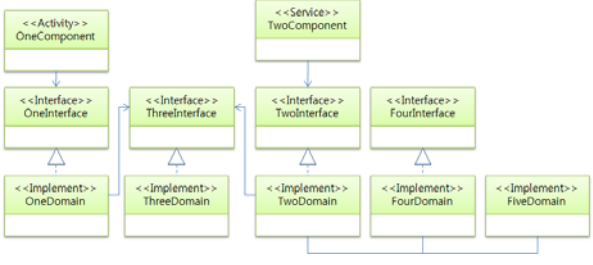


Fig. 4. Class diagram between each component and class

The components reference the interface classes for the implement classes and call the domain functions. And the implement classes that have association with components or other classes have to inherit the interface classes to make the domain functions. This approach reduces the direct relation between classes and create independent class model by separating each layers.

4 The Example Using the 4-Tire Design Pattern

In this chapter, we descript the design and implement method of Tracking App (Baugil) using the proposed design pattern. The Baugil App provides the service of a path information service with GPS and Map data based on android system. This application has the following features.

Information Sharing Service: This is web-based on/off-line information service that share user data and manage users information.

Path Tracking Service: This service provides path information of each course, tracking user location and the detail information of the significant points using GPS and Map data.

4.1 The 4-Tier Architecture

This Baugil Application architecture that is shown in Fig. 5 separate components and domain area and connect them with interface classes. This architecture consists of 7 executable components, 10 interface class and 12 implementation classes.

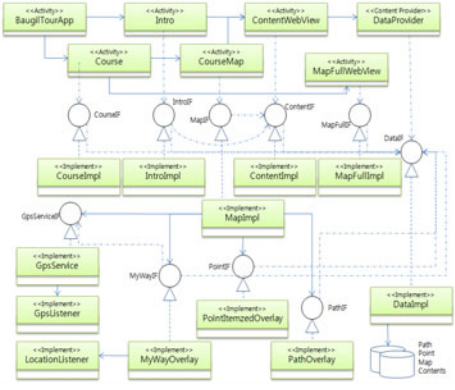


Fig. 5. The 4-Tier Architecture of the Baugil Application

The 7 executable components have the roles about allocation of resource, event management for the user interface with their life cycle and component management. And also, they control domain area with interface classes. The interface layer classes have the roles about call between methods in the related classes. It can be reduce dependency between classes. The domain layer that has implement classes has only their own functions for keeping a low coupling between classes.

4.2 The Relationship between Layers

We show a CourseMap structure to descript the relationship between layers in the Fig. 6.

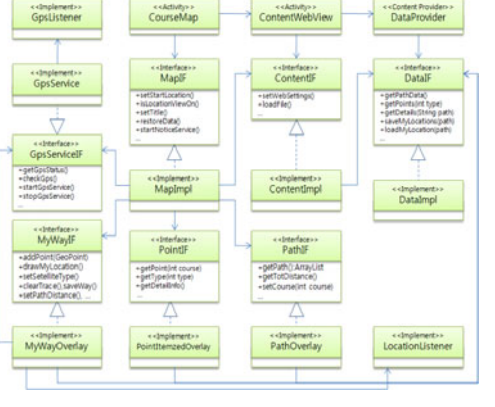


Fig. 6. The classes relation of the CourseMap

The CourseMap Activity class has MapIF interface class that have function prototype of the MapImpl implement class. And the Activity class communicates MapImpl class using MapIF interface. This Activity include load paths, notices, significant points and detail information and service tracking for user based on user current location when the start button click. This relation model shows relationship and related methods among the executable component, interface and implementation class.

4.3 The Implementation Result

The Implemented Baugil application using the proposed design pattern includes sharing information and user tracking service. Fig. 7 show the process steps for each components.

Fig. 7. The implementation result of the Baugil Application

The application has the sharing information service that provide on/off-line information service with web server and the tracking service that have 15 courses. the application draw paths based on latitude and longitude on the Google map and provide services to display start, end, notice, significant point, path length and user location information, to check the status of the satellite and to track user base on user current location. This application was implemented by separating to the each layer, included only their own features and connected each layer by using interface class independently. As a result, we could implement the application quickly and reliably.

5 The Conclusion

In this paper, we proposed the 4-Tire design pattern for the android application development. With the rapid growth of mobile environment, the demand of smart phones and the related applications has been increased rapidly. But because the application for the smart phone has special framework environment and life cycle, the general approach of the application design are difficult to reflect these characteristics.

So we proposed the approach to reduce the program complexity. In this approach, we separated an executable component of the android system and user domain area in the 4-tier layers. This approach could reduce the cost of the application, improve the ability of the maintenance and implement the application quickly and reliably.

References

1. Google Android Developer Guide, Android Framework and Application Fundamentals (2011), http://developer.android.com/guide/topics/fundamentals.html
2. Yang, J., Kim, E., Kim, N.: Android developer's guide, 3rd edn. (2010), http://www.kandroid.com/kandroid_book_3rd_edition.pdf
3. Jarvis, B., Gristock, S.P.: Extreme Programming (XP) SixSigma CMMI(2005), http://www.sei.cmu.edu/cmmi/
4. Wells, D.: Extreme Programming: A gentle introduction (2009), http://www.extremeprogramming.org
5. Jeong, Y.J., Lee, J.H., Shin, G.S.: Development Process of Mobile Application SW Based on Agile Methodology. In: Proceedings of the 10th International Conference on Advanced Communication Technology (2008)
6. Abrahamsson, P., Hanhineva, A., Hulkko, H., Ihme, T., Jaalinoga, J., Korkala, M., Koskela, J., Kyllonen, P., Salo, O.: Mobile-D: An Agile Approach for Mobile Application Development. In: Proceedings of 19th annual ACM SIGPLAN Conference on Object-Oriented Programming Systems, Languages and Applications (2004)
7. Natchetoi, Y., Kaufman, V., Shapiro, A.: Service-oriented architecture for mobile applications. In: Proceeding of the 1st International Workshop on Software Architectures and Mobility (2008)

A Study on the Power Divider of the Microstrip Antenna for Identification Friend or Foe Radar

Bong-Ki Jang* and Young-soon Lee**

Mobile Convergence Technology Center #307, 188, Synpung-dong, Gumi-city, Gyongbuk, Korea
jang815@hanmail.net, yslee@kumoh.ac.kr

Abstract. The final aim of the present study is the design on the power divider of the microstrip antenna for 1.03~1.09GHz IFF(Identification Friend or Foe) radar. First, in order to achieve the proper side lobe level, it is proposed nonuniform amplitude. For the purpose of supplying power to each array element, power divider which consists of 180 degree hybrid ring power divider and T-junction power divider is designed. Here 180 degree hybrid ring power divider is necessary for generating both sum and delta channel of IFF(Identification Friend or Foe) radar, while T-junction power divider is necessary to control power and decide radiation pattern. When compared with the required design specifications, it is expected that the present antenna can be applied directly to IFF radar.

Keywords: microstrip antenna, 180 degree hybrid ring power divider, T-junction power divider, IFF radar.

1 Introduction

Feed network has corporate type and series type. Series type has simple structure and is easy to implement, but it is difficult to feed same phase to all array elements. While corporate type has complicated structure and is easy to feed same phase.

For the purpose of supplying power to each array element, power divider which consists of 180 degree hybrid ring power divider and T-junction power divider is designed. Here 180 degree hybrid ring power divider is necessary for generating both sum and delta channel of IFF(Identification Friend or Foe) radar, while T-junction power divider is necessary to control power and decide radiation pattern.

In this paper, excitation coefficient will be calculated and array factor will be shown. By comparing uniform amplitude type and nonuniform amplitude type, it will be chosen Tchebyscheff array. To probe the validation, simulation results and measurement results are shown.

* Defense agency for Technology and Quality(Senior Researcher).
** Kumoh National University, Radio Communication Engineering(Professor).

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 204–214, 2011.

2 Excitation Coefficient

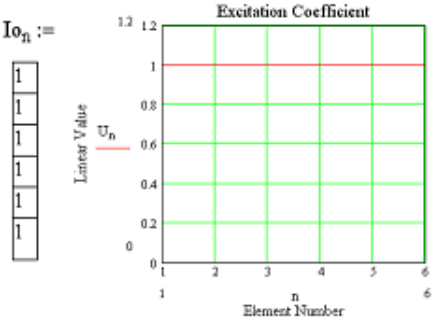
Array antenna is divided into uniform amplitude and nonuniform amplitude. Uniform amplitude is generally used communication system or satellite receiver system due to high gain. But uniform amplitude is not good on the focus of side lobe level. In case of IFF(Identification Friend or Foe) antenna, it is necessary –20dB side lobe level. Therefore, IFF antenna is proper nonuniform amplitude type like Binomial array, Tchebyscheff array and Cosine on Pedestal. Binomial array has very low side lobe and very broad main lobe and Cosine on Pedestal has low side lobe and high side lobe around the main lobe. Tchebyscheff array has sharp main lobe beamwidth and very low side lobe[1, 2]. Among various nonuniform amplitude type, Tchebyscheff array is the fittest type to the IFF radar antenna.

The normalized array of the uniform is given by[3]

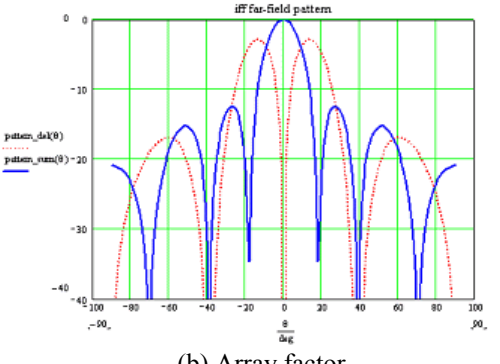
$$f(\theta) = \left| \frac{\sin(N \cdot \pi \cdot u(\theta))}{N \cdot \sin(\pi \cdot u(\theta))} \right| \tag{1}$$

Here, u(θ)=(d/λ)sin(θ), d : distance of array, N : number of array.

If the distance of array is 0.53λ and amplitude is uniform, excitation coefficient distribution and array factor show figure 1.



(a) Excitation coefficient distribution



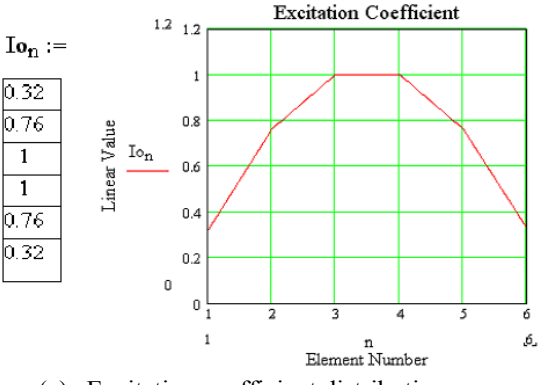
(b) Array factor

Fig. 1. Excitation coefficient distribution and Array factor(Uniform amplitude)

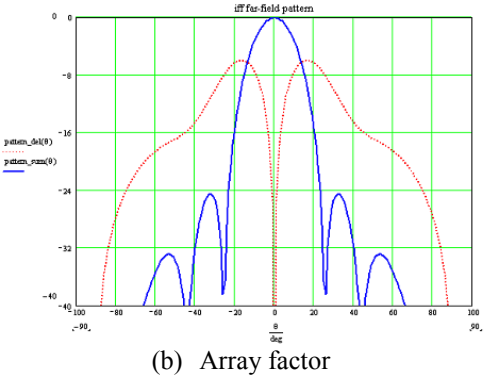
Tchebyscheff polynomial expression is given by[4]

$$T_n(x) = \begin{cases} (-1)^n \cosh(n \cosh^{-1}|x|), & x < -1 \\ \cos(n \cos^{-1} x), & -1 < x < 1 \\ \cosh(n \cosh^{-1} x), & x > 1 \end{cases} \quad (2)$$

For the purpose of high side lobe level(-21~-25dB), proper nonuniform amplitude show figure 2.



(a) Excitation coefficient distribution



(b) Array factor

Fig. 2. Excitation coefficient distribution and Array factor(nonuniform amplitude)

As shown Fig 2, when amplitude power is 0.32:0.76:1:1:0.76:0.32, half power beam width is about 20degree and side lobe level is about –21~-25dB. Power distribution of array 1~6 respectively shows Table 1.

Table 1. Power distribution of array

array	Linear Value	Power	Sub Sum(T-1)			Sub Sum(T-2)		
			Power	Normal Power	dB	Power	Normal Power	dB
1	0.32	0.102	0.680	0.405	-3.928	0.100	0.147	-8.327
2	0.76	0.578				0.580	0.853	-0.690
3	1.00	1.000	1.000	0.595	-2.253	N/A	N/A	N/A
4	1.00	1.000	1.000	0.595	-2.253	N/A	N/A	N/A
5	0.76	0.578	0.680	0.405	-3.928	0.580	0.853	-0.690
6	0.32	0.102				0.100	0.147	-8.327

3 Design of Power Divider

In this paper, Power divider is two types and show figure 3. The first is 180 hybrid ring power divider for the sum/delta channel. The second is T-junction power divider for the power control.

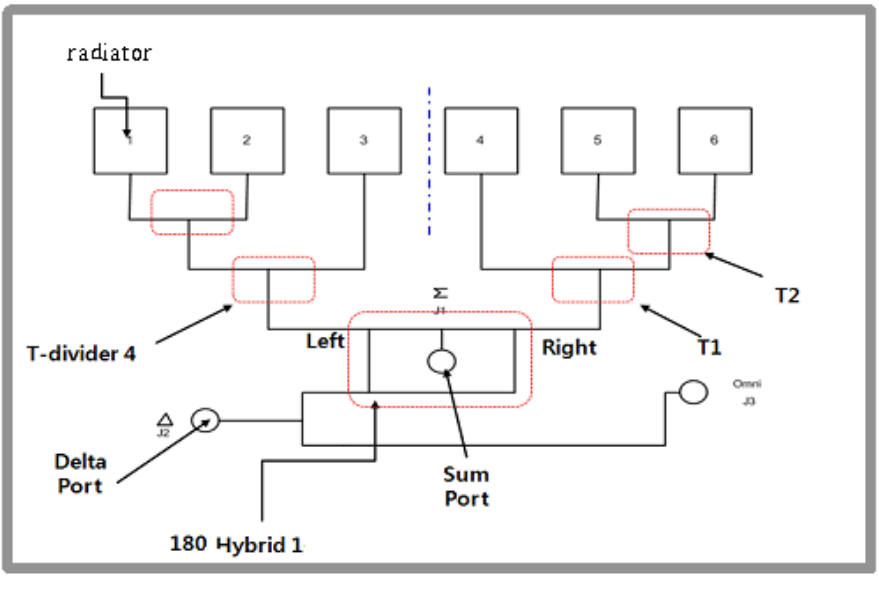


Fig. 3. Structure of Power divider

3.1 180 Degree Hybrid Ring Power Divider

180 hybrid ring power divider consists of Sum Port, Delta Port, Output1 Port, Output2 Port and show Fig 4. Sum Port is combined by Output1 Port and Output2 Port. Delta Port is subtracted by Output1 Port and Output2 Port. Finally sum pattern and delta pattern is implemented by180 hybrid ring power divider.

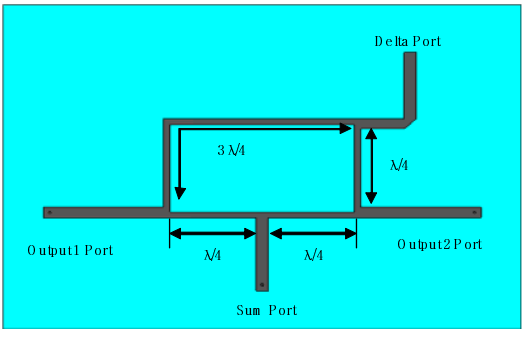


Fig. 4. Design of 180 hybrid ring power divider

In 180 hybrid ring power divider, width(w) of line can be designed by power ratio and impedance transformation. If sum port defines P1, output1 port defines P2 and output2 port defines P3, parameter show Fig 5.

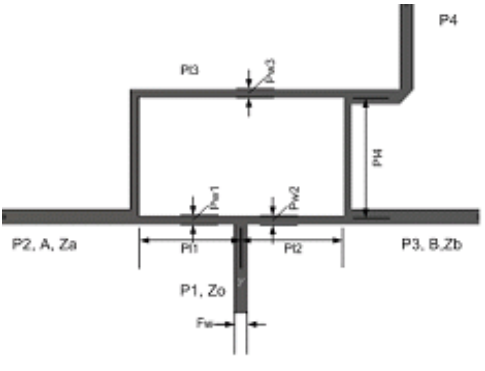


Fig. 5. Parameter of 180 hybrid ring power divider

By using MWS CST analysis program, Fig 6 shows return loss when width changes. As knows from Fig 6, width of input feeder line(Fw) 5.63mm is the fittest value.

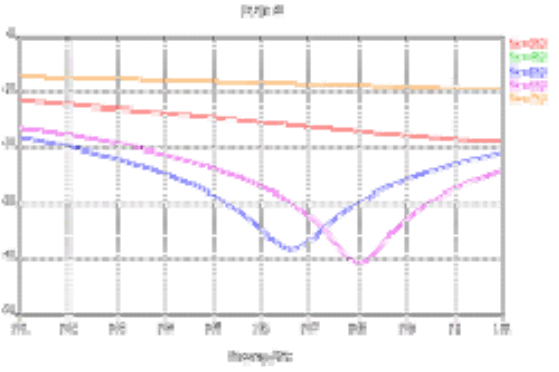


Fig. 6. Return loss analysis result

Fw is width of input feeder line, Pw1~Pw3 is width of dividing input feeder line, and Pl1~Pl4 is length of feeder line. Detail parameters are Table 2.

Table 2. Parameter of 180 hybrid ring power divider

Parameter	Dimension(mm)
Width of input feeder line(Fw)	5.63mm
Width of Port2 feeder line(Pw1)	3.08mm
Width of Port3 feeder line(Pw2)	3.08mm
Length of the P1-P2 feeder line(Pl1)	42.63mm
Length of the P1-P3 feeder line(Pl2)	42.63mm
Length of the P4-P2 feeder line(Pl3)	127.89mm
Length of the P4-P3 feeder line(Pl4)	42.63mm

Fig 7 ~ Fig 11 is return loss, power dividing ratio and phase value. From analysis, P1(Sum) return loss is –21.4dB, P4(Delta) return loss is –15.8dB and satisfy the design specification –13.97dB(VSWR 1.5:1). Power ratio is P1-P2 and P1-P3 is respectively -2.929, -3.067, P4-P2 and P4-P3 is respectively –3.446, -3.252 and is almost satisfy the design specification –3dB(half power). On the basis of P1, Port phase difference is 0.74degree(P2, P3 respectively phase is 56.88degree and 56.14degre). On the basis of P4, Port phase difference is 0.756degree(P2, P3 respectively phase is 5.356degree and –175.4degree).

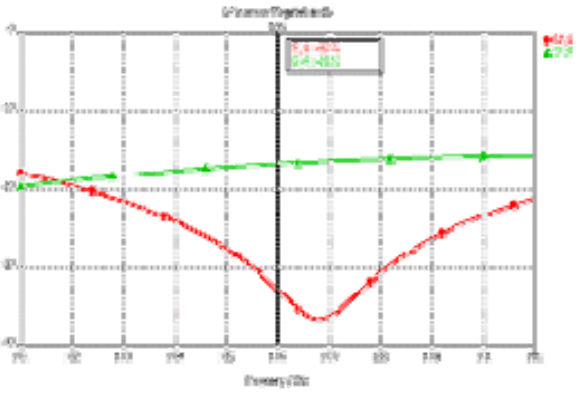


Fig. 7. Return loss analysis result

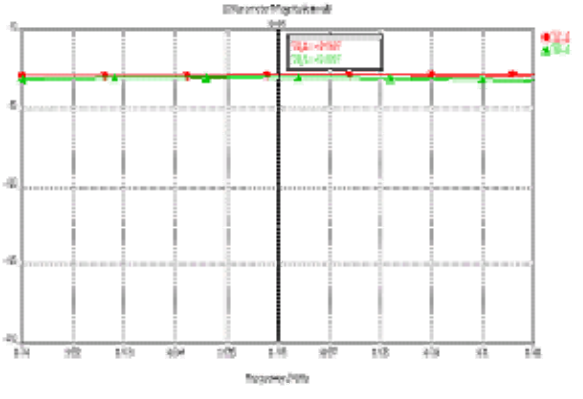


Fig. 8. Power divider(Sum channel)

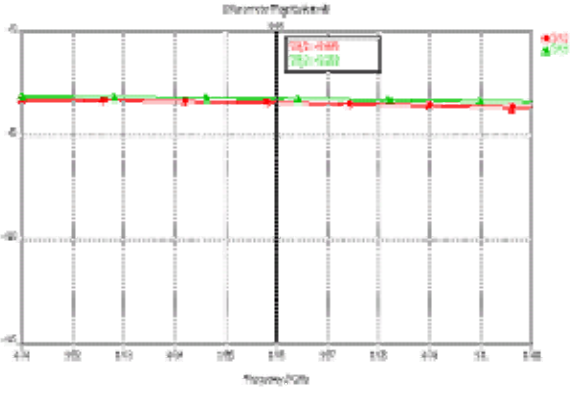


Fig. 9. Power divider(Delta channel)

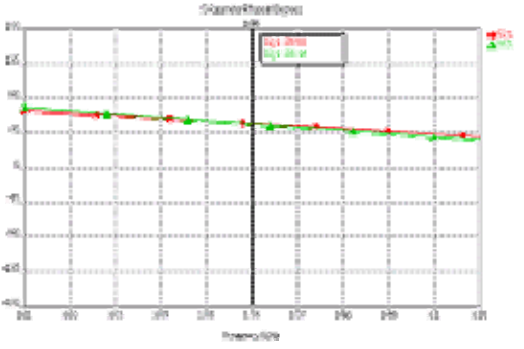


Fig. 10. Phase characteristic(Sum channel)

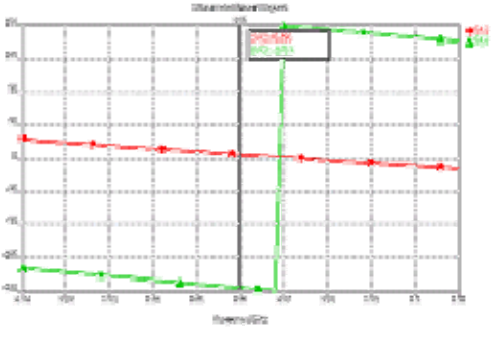


Fig. 11. Phase characteristic (Delta channel)

3.2 T-junction Power Divider

T-junction power divider consists of Input Port, Output1 Port, Output2 Port and show Fig 12. Sum Port is combined by Output1 Port and Output2 Port. Delta Port is subtracted by Output1 Port and Output2 Port. Finally sum pattern and delta pattern is implemented by180 hybrid ring power divider.

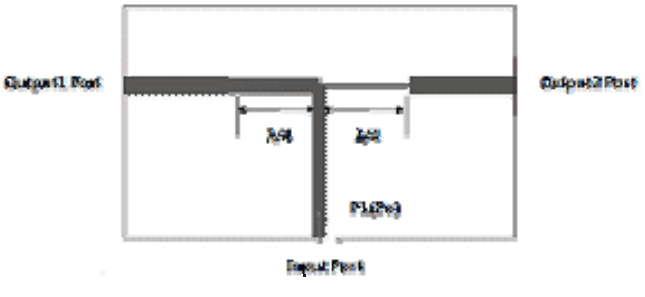


Fig. 12. Design of T-junction power divider

In T-junction power divider, width(w) of line can be designed by power ratio and impedance transformation. If input port defines P1, output1 port defines P2 and output2 port defines P3, parameter show Fig 13, Fig 14.

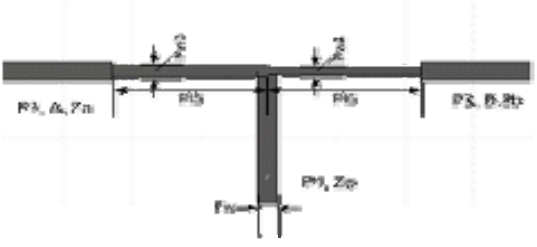


Fig. 13. Parameter T-junction(T-1) power divider

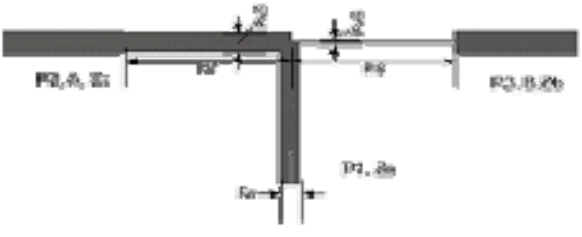


Fig. 14. Parameter T-junction(T-2) power divider

By using MWS CST analysis program, Fig 15 shows return loss when width changes. As knows from Fig 15, width of input feeder line(Fw) 5.63mm is the fittest value.

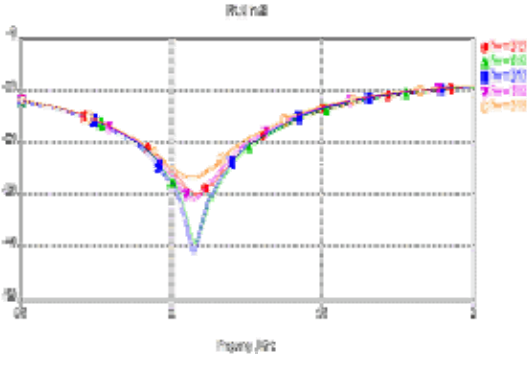


Fig. 15. Return loss analysis result

Fw is width of input feeder line, Pw3~Pw6 is width of dividing input feeder line, and Pl5~Pl8 is length of feeder line. Detail parameters are Table 3.

Table 3. Parameter of T-junction power divider

Parameter	Dimension(mm)
Width of input feeder line(Fw)(Fw)	5.63mm
Width of T-1 Port2 feeder line(Pw3)	3.69mm
Width of T-1 Port3 feeder line(Pw4)	2.44mm
Width of T-2 Port2 feeder line(Pw5)	4.96mm
Width of T-2 Port3 feeder line(Pw6)	0.49mm
Length of T-1 Port2 feeder line(Pl5)	42.63mm
Length of T-1 Port3 feeder line(Pl6)	42.63mm
Length of T-2 Port2 feeder line(Pl7)	127.89mm
Length of T-2 Port3 feeder line(Pl8)	42.63mm

4 Implementation of Power Divider

Final implementation show Fig 16. Table 4 is design simulation result and implementation test result.

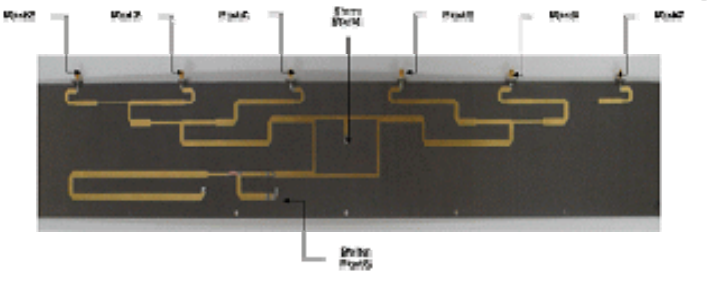


Fig. 16. Return loss analysis result

Table 4. Simulation result and Implementation result

Port	Power distribution(dB)<1.03GHz,∑>		
	Simulation (CST MWS)	Implementation (Test)	difference
2	−15.38	−16.14	0.76
3	−8.133	−8.306	0.173
4	−5.32	−5.274	0.046
5	−5.306	−5.249	0.057
6	−8.124	−8.515	0.391
7	−15.35	−16.20	0.85

5 Conclusion

For the design on the power divider of the microstrip antenna for 1.03~1.09GHz IFF(Identification Friend or Foe) radar, it is proposed nonuniform amplitude. Amplitude power is 0.32:0.76;1:1:0.76:0.32, half power beam width is about 20degree and side lobe level is about –21~-25dB.

Power divider consists of 180degree hybrid ring power divider and T-junction power divider. Respectively implementation results satisfy design specification.

From this paper, it is expected that the present antenna can be applied directly to IFF radar.

References

1. Bailey, A.E. (ed.): Microwave Measurement, Engineering. Peter Peregrinus, London (1985)
2. Collin, R.E.: Foundations for microwave Engineering, 2nd edn. McGraw-Hil, NY (1992)
3. Wilkinson, E.J.: An N-way Hybrid Power Divider. IEEE Trans. Microwave Theory and Tech. 8, 116–118 (1960)
4. Pon, C.Y.: Hybrid Ring Directional Couplers for Arbitrary Power Division. IEEE Trans. Microwave Theory and Tech. 9, 529–535 (2006)

An Approach to Access the Distributed Data Based on the Multi-Agent System for Interoperability

Youn-Gyou Kook, Joon Lee, Min-Woo Park, Jae-Soo Kim, and Ki-Seok Choi*

Dept. of NTIS (National Science & Technology Information Service)
KISTI (Korea Institute of Science and Technology Information)
Daejeon, Korea,
{ykkook,rjlee98,choi}@kisti.re.kr

Abstract. In this paper, we present an approach to access the distributed data for interoperability in the distributed environments based on the multi-agent system that is designed on the proposed structure of multi-agent by FIPA(IEEE Foundation for Intelligent Physical Agents). This multi-agent system is light weight platform at the legacy system to be without external influence. To build up a domain that is consists of the legacy systems, we deploy the multi-agent platform at that system to perform the goal of the domain. The proposed approach can build up the various multi-domain configurations have the goal that must be performed in the distributed environments. This reduced the cost of building the infrastructure to interoperate the distributed data and expanded the scope of using that.

Keywords: Multi-Agent System, Legacy System, Distributed Data, Interoperability, Data Access Agent.

1 Introduction

It is important to research and make use of interoperating the distributed data among the legacy systems that are autonomous and independent without external control. There are the methods of interoperating the data which are data integration based on scheme integration of the distributed data and data connection through data migration and transformation among the legacy systems. The scheme integration is needed to be advanced legacy systems for the distributed data integration with large cost and time, whereas the data connection is requested for the least cost and time of modifying the legacy systems to guarantee autonomous and independent operating. So it uses the approach of data connection based on the infrastructures which are client/server structure, web-service, data-grid system and etc [6]. The mainly role of these is transferred data to the server that has the process of transformation and cleansing to be suitable for interoperating the distributed data. And it can retransfer and reload the processed data to the legacy systems which are involved in the interoperating. But it is

*Corresponding Author.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 215–222, 2011.

difficult to solve the problems that are maintaining the coherence and managing the quality of data.

So we propose the approach to access the distributed data based on the multi-agent system for interoperability to adhere to distributed environments that is minimized an influence without external control. Multi-agent system cooperates with other agents and some application programs and has the specific goal to solve a problem within a domain. However, it is built up the domain in the centrally multi-agent platform for interoperating the distributed data to perform that goal, because this topology is centralized by Master of multi-agent framework. Therefore we propose the advanced the topology of multi-agent to access the distributed data among the legacy systems. The proposed approach is built up the various multi-domain configurations based on the light-weight multi-agent platform.

In this paper, we present the methods of building up the various multi-domain based on the multi-agent system that is followed by the base concept of FIPA(IEEE Foundation for Intelligent Physical Agents) [5] and the approach of accessing that data through the methods. We designed and implemented the multi-agent platform that is light-weight included in AMS(Agent Management System). That platform is deployable at the legacy systems to be without external control. The proposed approach can be reduced the cost and time of building the infrastructure of interoperating the distributed data and minimized the external influence to guarantee the autonomy and the independent operating. And this multi-agent system can perform the various goals.

2 Multi-Agent System

Software Agent is a program that is defined as ‘a system to be able to cooperate with the other agent and users for performing the specific goals in place of user-performed that’. This agent has the various characters at the specific goals as follows that: Autonomy, Intelligence, Mobility, Sociality, Reactivity, Veracity and so on[1, 2]. And Multi-Agent System is defined as follows that: an agent to process tasks by cooperating with the various multi-agent and application programs to perform the specific goal. That is provided the framework can be executed the various multi-agents, the method of building up a domain included that agents and the channel to exchange messages between agents. In particular, MAF(multi-agent framework) is the centralized topology in use of Coordination Agent to exchange messages. This is the Agent Platform proposed by FIPA that is widely known as MAF [3, 4, 5]. Figure1 is shown as that.

The centralized topology of multi-agent system is consists of Master and Slave. Two modules exchange messages to communicate between them. Master module is included ACC(Agent Communication Channel), AMS(Agent Management System), DF(Directory Facilitator), ANS(Agent Name Server) and Slave module is included the various multi-agent and application programs. This is designed to be suitable for performing a specific goal which that multi-agent system can execute the tasks.

An agent manager in a domain based on multi-agent platform manages the various multi-agents. Figure2 is shown as that. The domain1 is consists of 3 legacy systems, has one agent platform and 3 agents. An agent manager of agent platform manages the agent of server1, server2 and server3 shown as figure2.

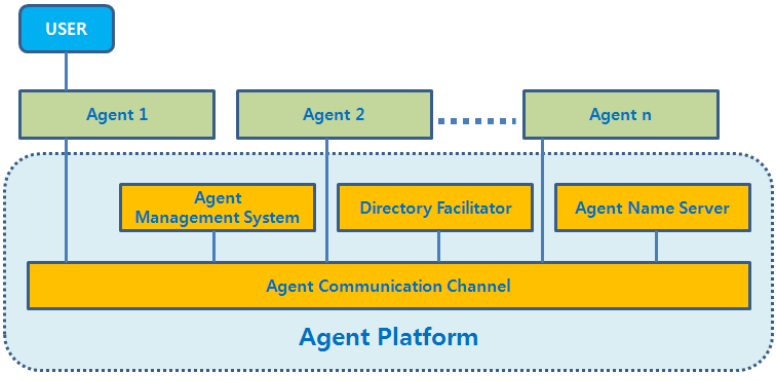


Fig. 1. The steps of data quality management

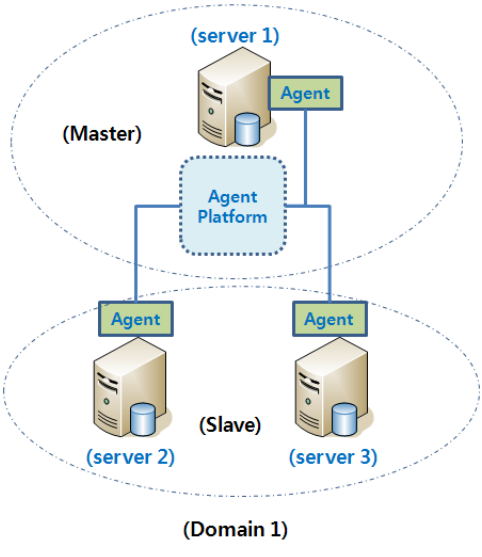


Fig. 2. The steps of data quality management

However, the distributed topology of multi-agent system in the proposed approach can build up the multi-domain configurations for performing the various multi-goals. Each domain includes the various multi-agents to perform the specific goal and the multi-agents belong to the multi-domains. The multi-agent system at the legacy systems execute the both roles of master and slave.

3 The Multi-Agent System on the Distributed Topology

It is possible for the multi-agent system build up a domain included the various multi-agents for performing the specific goal. However a domain of the centralized topology can be execute only one goal to execute tasks that are carried out by the various multi-agent and application programs. It is difficult to extend the range of domain for interoperating the distributed data in wide distributed environments. But, the distributed topology of the multi-agent system is available to build up the various multi-domains which have their own goals. This topology has the multi-domains as shown as figure 3.

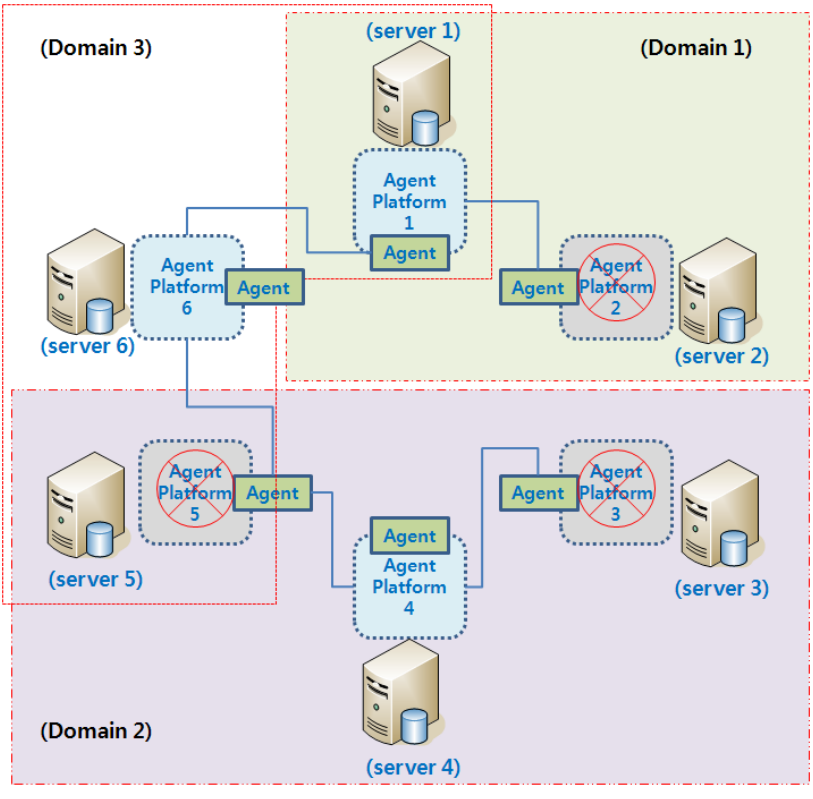


Fig. 3. The steps of data quality management

In figure 3, there are 3 domains that are consisted of the multi-agent systems and 6 legacy systems as shown as servers. Each domain is operated in autonomous and independent to perform the goal and is available to cooperate between them. All of the legacy systems have the agent platform included AMS, but a part of the AMS is usable for building up a domain. Let D is a set of domain lists, D = {$domain_1$, $domain_2$, …, $domain_m$, $domain_n$}, L is a set of legacy systems, L = {$server_1$, $server_2$, …, $server_m$, $server_n$}. One server has one agent platform which is multi-agent system and the various

multi-agents. The $server_k$ is available to be intersected in the over two domains for building up the domain.

$Domain_1$ is consists of $server_1$ and $server_2$, $Domain_2$ is consists of $server_3$, $server_4$ and $server_5$ and $Domain_3$ consists of $server_1$, $server_5$ and $server_6$. Agent $platform_1$, agent $platform_4$, agent $platform_6$ is available for building up domains, but agent $platform_2$, agent $platform_3$ and agent $platform_5$ is not available. But all of the multi-agents is available for interoperating in the distributed environments. AMS of agent $platform_1$ at $server_1$ manages the agent of $server_1$ and $server_2$. AMSs of agent $platform_4$ at $server_4$ and agent $platform_6$ at $server_6$ manage agents like the preceding. In particular, the agent of $server_1$ in $domain_1$ and $domain_3$ is available for executing their each task by requesting of each agent platform. All of the agents return the results of executing task to the AMS in which agent platform request to perform the goal.

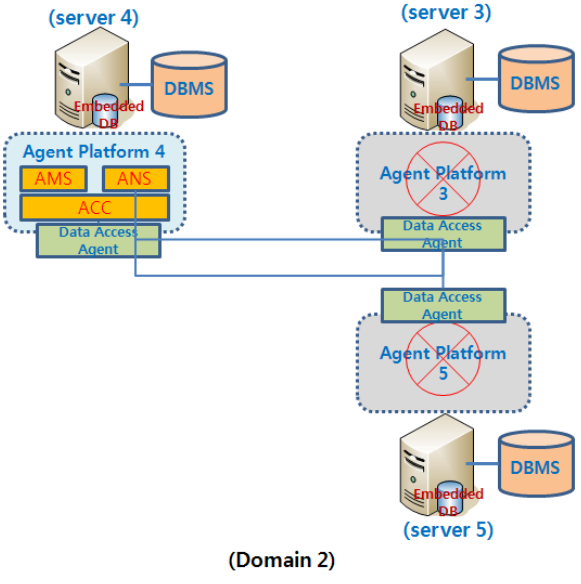


Fig. 4. The steps of data quality management

Especially, the distributed topology based on the multi-agent system is suitable for interoperating data in the distributed environments without external influences. It is guaranteed the autonomy and independent operating of the legacy system. Figure 4 shows the use of multi-agent system to interoperate the distributed data in $domain_2$. Each server has their own embedded DB and DBMS that is managed by self-operating. The agent platform of server includes with ACC, AMS, ANS and DAA(Data Access Agent). The DAA accesses to DBMS of legacy systems by requesting of AMS. The DAAs of $server_3$, $server_4$ and $server_5$ register to the ANS of agent $platform_4$. Of course, agent $platform_3$ and agent $platform_5$ is not available.

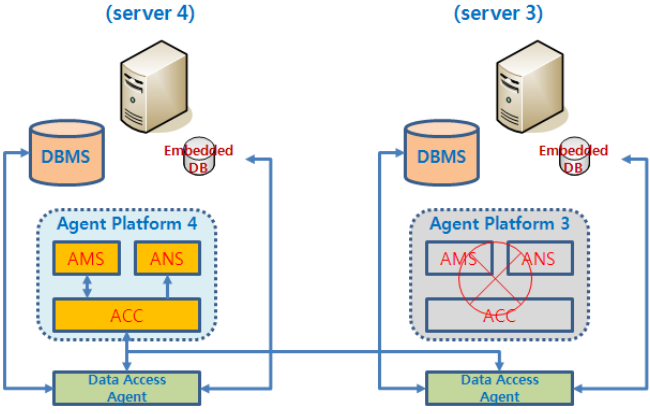


Fig. 5. The steps of data quality management

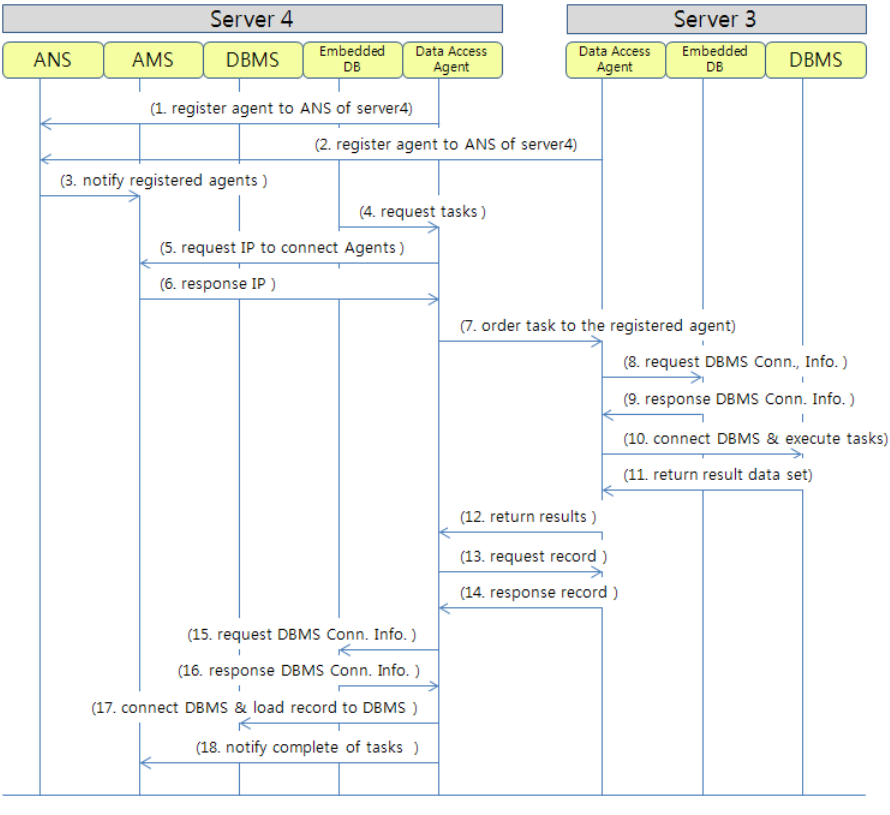


Fig. 6. The steps of data quality management

The feature of connecting with communication channel to exchange a message for interoperating the local DBMS in two legacy systems is shown as figure 5. First, the DAAs of $server_3$ and $server_4$ register to the ANS of agent $platform_4$ through ACC. The ANS notifies the registered agent lists to the AMS of agent $platform_4$. The DAA of $server_4$ receives the request of executing tasks from embedded DB included fetching task information by the AMS. Second, the DAA of $server_4$ requests IP of the other agents to connect with the agents to the AMS and receive that information. Third, the agent of $server_4$ requests to access the DBMS to the DAA of $server_3$. Forth, the DAA of $server_3$ requests the DBMS connection information to the embedded DB of $server_3$ and receives that information. And that accesses to the DBMS of $server_3$ to execute tasks and returns the results. Fifth, the DAA of $server_3$ returns the results to the DAA of $server_4$. Sixth, the DAA of $server_4$ requests the record of result set and receives that. After the DAA of $server_4$ receives the DBMS information of $server_4$, that DAA loads the record is received from $server_3$ to the DBMS of $server_3$. Finally, the DAA of $server_4$ notifies the complete sign to the AMS of $server_3$. These steps are shown as figure 6.

4 Conclusion

There are the methods of interoperating the data which are data integration based on scheme integration of the distributed data and data connection through data migration and transformation among the legacy systems. It is important to interoperate the distributed data without external control in the distributed environments, because the legacy systems are guaranteed the autonomy and independent operating. The multi-agent system for interoperability to adhere to distributed environments that is minimized an influence without external control. Multi-agent system cooperates with other agents and some application programs and has the specific goal to solve a problem within a domain.

Therefore we propose the advanced the topology of multi-agent to access the distributed data among the legacy systems. We described the approach is built up the various multi-domain configurations based on the light-weight multi-agent platform and is accessed to the DBMS of legacy systems based on the multi-agent system.

So we presented the methods of building up the various multi-domain based on the multi-agent system and the approach of accessing that data through the methods. We designed and implemented the light-weight multi-agent platform. That platform at the legacy systems is minimized the influence of external control. This approach can be reduced the cost and time of building the infrastructure of interoperating the distributed data and provided the infrastructure of performing the various goals.

References

1. Nwana, H.S.: Software Agents: An Overview. Knowledge Engineering Review 11(3), 1–40 (1996)
2. Wooldgridge, M.J., Jennings, N.R.: Agent Theories, Architectures and Languages: A Survey. In: Proceedings of ECAI 1994 Workshop on Agent Theories Architectures and Languages, pp. 1–39 (1994)
3. Obrien, P.: Agent Management. In: FIPA 1997 Draft Sepcification: Part1, p. 38 (1997)
4. Hull, R.: Managing semantic heterogenceity in database: a theoretical prospective. In: Proceedings of the 16th, ACM SIGACT-SIGMOD-SIGART symposium on Principles of Database Systems, pp. 51–61 (1997)
5. IEEE Foundation for Intelligent Physical Agents, `http://www.fipa.org`
6. Aloisio, G., Cafaro, M., Fiore, S., Mirto, M.: The Grid-DBMS: Towards Dynamic Data Management in Grid Environments. In: Information Technology: Coding and Computing (ITCC 2005), vol. 2, pp. 199–204. IEEE (April 2005)

Design and Implementation of a Remote Control for IPTV with Sensors

Jae Ha Song[1], Woo Yeol Kim[2], Hyun Seung Son[2], Junbeom Yoo[3], Jae Seung Kim[2], Robert Young Chul Kim[2], and Jung Hun Oh[1]

[1] Deptartment of Embedded System, Baron Co. Ltd., Siheung-si, Korea
{songjh3218,119OJH}@paran.com
[2] Dept. of CIC(Computer and Information Communication), Hongik University, Jochiwon, 339-701, Korea
{john,son,jskim,bob}@selab.hongik.ac.kr
[3] Department of Computer Science and Engineering, Konkuk University, Seoul, 143-701, Korea
jbyoo@konkuk.ac.kr

Abstract. Existing TV remote controls using relative coordinates have some difficulties in selecting various types of IPTV content. This paper proposes an absolute coordinates based remote control to make the selection of IPTV channels more convenient. It uses a laser pointer to point at an IPTV, a camera mounted the remote control to locate the spot pointed at, and control software to calculate the absolute coordinates of the spot pointed to on the screen.

Keywords: IPTV, Remote Control, Absolute Coordinates.

1 Introduction

Internet Protocol television (IPTV) [1] is a system through which digital television service is delivered using internet protocol, while other TVs use traditional radio frequency broadcast or cable television formats. IPTV provides various services via two-way communications, such as VOD (Video On Demand), shopping, online video games, singing rooms and chatting as well as TV channels [2-4]. It needs a set-top box for the IP based services and an advanced remote control to utilize the services easily and efficiently [5].

While the latest set-top box technologies have offered a number of interactive services, the remote controls have found some difficulties in keeping up with them. They just provide additional arrows buttons for moving a cursor, and it makes the interactive selection of contents difficult [6-8]. When pressing an arrow button, the cursor does not move directly to the point or contents on the screen which a user wants, but moves to the content next to the current one. It creates the need for a number of button presses and inconvenience.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 223–228, 2011.

The paper is organized as follows. Section 2 discusses related work on existing IPTV remote controls. Sections 3 and 4 explain the design and implementation of the proposed remote control, respectively.. Section 6 concludes the paper.

2 Remote Control for IPTV

IPTV remote controls can be divided into 3 types: remote controls with a number of buttons; multi-function remote controls with a touch screen; and motion remote controls with internal motion sensors.

A button type remote control, such as the hanaTV remote [9], is easy to implement since it is the same as the existing remote controls. It however has some drawbacks. It is difficult to update as new functions, contents and channels are added. It is also too slow for users to move the cursor into the location where he wants, since he has to push arrow buttons several times.

Touch screen type remote controls were proposed by Caviar [10] and Harmony1000 [11]. Caviar used a 3.5 inch touch screen, but it was too expensive for use as an IPTV remote control. Logitech Harmony1000 is an integrated multi-media remote control with a 3.5 inch touch screen like the Caviar. The difference is that the Harmony can be updated through the Internet. The updating procedure however is complicated for general users to manage. It uses a joy-stick for cursor movements instead of buttons.

A couple gyro sensor installed motion remote controls are the Wii remote [12] and the Magic motion remote control [13]. The Wii remote for the Wii game console is a type that is controlled with the use of an acceleration sensor. The acceleration sensor is installed internally so that the control is possible with different motions. For example, circling and crossing movements are translated into the execution of selecting and canceling. The Wii remote is mostly used for gaming content. The Magic motion remote control is for Xcanvas by LG Electronics. Like the Wii remote, the Magic motion remote control with an acceleration sensor can control screen shifts and movements. However, both products rely on a relative coordinate system which lacks direct selection from the objects on the screen.

3 Design of Remote Control

As seen in Figure 1, the system to control the contents of an IPTV consists of a TV set, an IPTV set-top box, a wireless receiving circuit and a camera-type remote control. As for the way of operating the system, when placed at a spot aimed by the remote control, the wireless transmitter will convey the coordinate information to the set-top box. The IPTV software installed inside the set-top box receives the location coordinate aimed at by the remote control and marks the cursor on the TV screen. The set-top box will be perceived to be like a mouse, freely selecting the contents on the TV screen. In this paper, a remote control, the core out of the overall system composition, is designed and implemented.

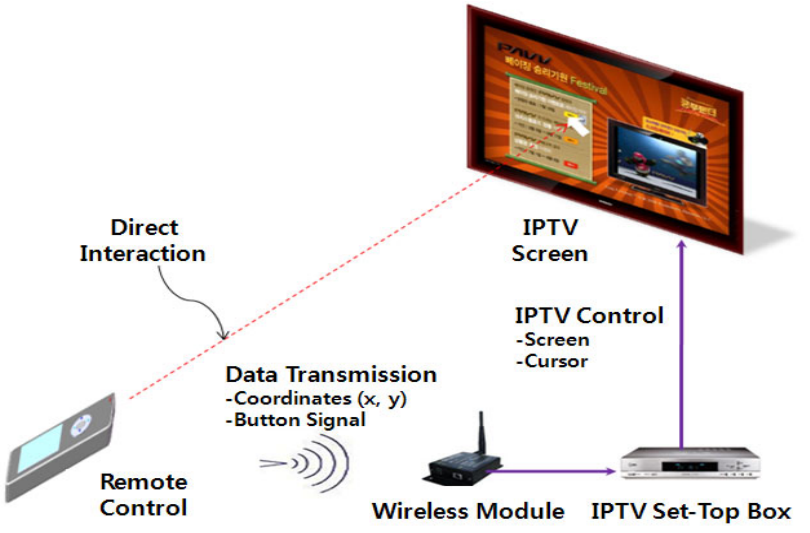


Fig. 1. The system architecture of the IPTV remote control

The camera-typed remote control consists of a CCD camera, an A/D convertor, FPGA. MCU and a wireless transmitter, as shown in Figure 2.

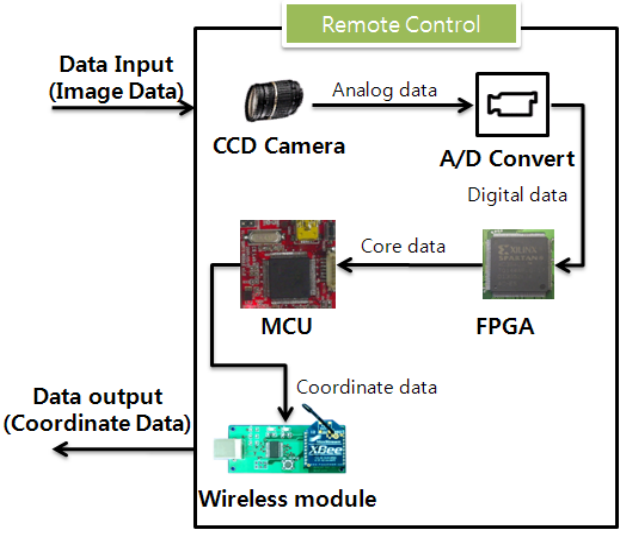


Fig. 2. H/W architecture for a remote control based on a camera

The image signals produced are generally analog signals, but in the case of a small-sized camera built into a cell phone, they are digital image signals. When choosing a digital camera outputting digital image signals, it is not necessary to design an additional A/D transformation circuit. On the other hand, a CCD camera is usually

analog, so it needs an A/D converter. In the case that a camera with a 1280 × 1024 resolution outputs images at 15 times per second, it is required to choose a FPGA able to process 8 bit data at over 30 MHz speed.

The FPGA plays a role in receiving the camera image signal data and extracting the core data. The MCU calculates the point coordinate aimed at and transmits the coordinate and button signals calculated to the set-top box through the wireless transmitting circuit. The remote control proposed can be completed by designing and assembling hardware and firmware, such as a camera, a lens, FPGA, MCU, and a wireless transmitting circuit in a proper case.

4 Implementation of the Remote Control

To implement the remote control, the IR filter is built in the front of the lens as shown in Figure 3 so that only images in the infrared-ray domain can be accepted as input into the camera in the front of the remote control. Since both interior and exterior images of the TV set correspond to the domain of visible rays, it will be easy and simple to extract the location of the TV screen in the case of using the infrared-ray LEDs as particular marks.

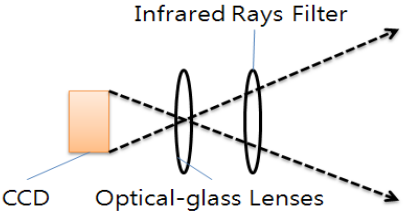


Fig. 3. Optical Composition of the Camera to Perceive Particular Marks

The wavelength of visible rays is defined to below 700 nm, but in other cases, it can be defined to be 750 nm or even below 800 nm. For example, if the wavelength of the infrared-ray LED is 770 nm and the IR filter passes wavelengths over 750 nm, the luminescence domain of the infrared-ray LEDs should be perceived to be brighter. In an actual circumstance, there exist visible rays over 750 nm, so there may exist a low brightness illuminated domain in addition to the domain of the infrared-ray LEDs. Even though it is less bright than the luminescence domain of the infrared-ray LEDs, any domains with a little illumination can be obstacle factors in the process of extracting only the domain of the infrared-ray LEDs. To prevent the visible rays from going into other domains as much as possible except for the particular marked domains, it is required to use infrared-ray LEDs with wavelengths over 800 nm and set the wavelength of the IR filter to over 800 nm. Since a camera mostly perceives the domain of visible rays and infrared wavelengths below 1,000 nm, the wavelengths of the infrared-ray LEDs and IR filter need to be set to over 800 nm and below 1,000 nm. Wavelengths of 850 nm for the infrared-ray LED and of 800 nm for the IR filter should be chosen for the proposed remote control.

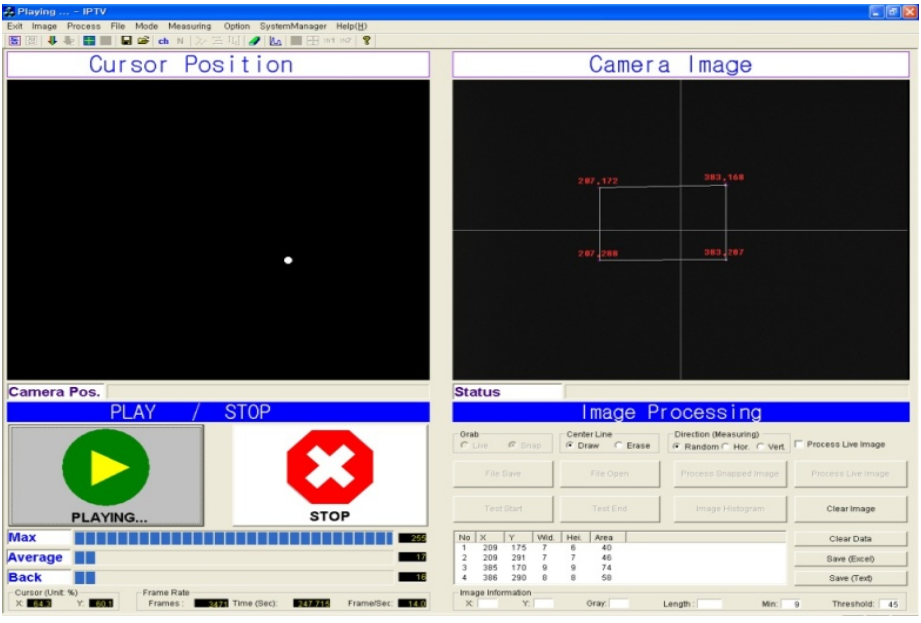


Fig. 4. GUI of the software for IPTV

Figure 4 shows the prototype implementing the software designed. On the left side is marked the coordinate of the screen cursor on the screen, and on the right side is the image information input from the camera on the screen.

5 Conclusion

IPTV (Internet Protocol Television) provides a variety of two-way services, such as a VOD service, a TV home shopping service, on-line games, a singing-room service and an internet chatting service, in addition to the existing TV channels. With existing remote controls it is difficult to easily select among the hundreds of types of content available through IPTV.

To solve this problem of the existing remote control technology, this research designed and implemented a new remote control for IPTV based on absolute coordinates. The remote control methodology is to install four infrared-ray LEDs on the TV set and to calculate coordinates through a real-time analysis of the images taken by the camera. For this process, an algorithm was developed to create coordinates with the input images from the camera by using trigonometric functions and an equation to calculate the distance.

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-2011-(C1090-1131-0008)))

References

1. Yang, X., Xiaojiang, D., Jingyuan, Z., Fei Hu, G.: Internet Protocol Television (IPTV): The Killer Application for the Next-Generation Internet. IEEE Communications Magazine 45(11), 126–134 (2007)
2. Kerpez, K., Waring, D., Lapiotis, G., Lyles, J.B., Vaidyanathan, R.: IPTV service assurance. IEEE Communications Magazine 44(9), 166–172 (2006)
3. Ji-Eun, L., Minsoo, S.: The role of public services in the convergence era: IPTV case. Info. 12(1), 39–53 (2010)
4. Montpetit, M.J., Klym, N., Mirlacher, T.: The future of IPTV: Adding social networking and mobility. In: 10th International Conference on Telecommunications 2009 (ConTEL 2009), pp. 405–409 (2009)
5. Ortiz Jr., S.: Phone Companies Get into the TV Business. Computer 39(10), 12–15 (2006)
6. Dan-Hee, Y.: IPTV Service and Improvement in Korea. Review of Korean Society for Internet Information 8(1), 23–28 (2007)
7. Enns, N.R.N., MacKenzie, I.S.: Touchpad-based remote control devices. In: Proceeding CHI 1998 Conference Summary on Human Factors in Computing Systems (1998)
8. Jana, R., Yih-Farn, C., Gibbon, D.C., Yennum, H., Jora, S., Murray, J., Wei, B.: Clicker - An IPTV Remote Control in Your Cell Phone. In: 2007 IEEE International Conference on Multimedia and Expo, pp. 1055–1058 (2007)
9. hana TV, `http://www.hanatv.co.kr`
10. utas, `http://www.utas.co.kr`
11. Harmony1000, `http://www.logitech.com`
12. Wii remote, `http://www.nintendo.com/`
13. Magic motion remote control, `http://www.xcanvas.co.kr/`

The End-to-End Reliability Algorithms Based on the Location Information and Implicit ACK in Delay Tolerant Mobile Networks

Doo-Ok Seo* and Dong-Ho Lee**

Dept. of Computer Science, Kwangwoon University, Seoul, Korea
clickseo@kw.ac.kr, dhlee@kw.ac.kr

Abstract. While portable and wireless devices are pouring, a new network technology is needed as a breakthrough. The new network technology features large delays, intermittent connectivity, and absence of an end-to-end path from sources to destinations. A network which has one of those characteristics is called DTNs(Delay Tolerant Networks). The main 4 routing methods have been researched so far in extreme environment. In this paper, we look into the reliability of DTMNs(Delay Tolerant Mobile Networks) in several different situations, and propose an algorithm that selects a positive routine by sending the only information of its position when making a connection to a detected node. We simulate the proposed algorithm here in DTN using ONE simulator. As a result, it shows that the algorithm reduces the number of message transmission each node.

Keywords: Delay Tolerant Networks, DTMNs, Mobility, Reliability, ONE simulator.

1 Introduction

Wireless devices such as laptop, PDA and mobile phones are increasing explosively. User's mobile device dependency is also increasing and it is required to be connected whenever, wherever users need to do. These requirements are also applied in extreme network environments. For example, network environments such as satellite network, earth to planet communication, military network, isolated remote village and disaster rescue are more prominent with recent nature disaster events. Therefore, it is required to communicate in extreme network environments.

The appearance of new network environments force us to take new technology adventure such as network split, large delay, intermittent connectivity, high link error rate and network environments between heterogeneous devices. Therefore new hypotheses are required such as large delay, intermittent connectivity and the most importantly absence of path between source and destination.

New adventures and hypotheses accelerates researches in extreme mobile network. Mobile Ad Hoc NETworks(MANETs) researchers focused on routing issues in order to solve mobility problem[1-3]. However, they only considered end-to-end path

* Ph.D. Student of Computer Science, Kwang-Woon University.
** Professor Deot. of Computer Science, Kwang-Woon University.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 229–239, 2011.

absence between source and destination without considering other extreme environments. Another research group focused on addressing in order to solve path absence issue and introduced Delay Tolerant Networks(DTNs) structure focusing on routing and message transmission issue in extreme environments[4-7].

In this paper, we will introduce 3 end-to-end reliability research for a specific DTNs structure, Delay Tolerant Mobile Networks(DTMNs), intermittent connectivity mobile network in large area[8]. In chapter 2, we will introduce relevant research and DTMNs summary. In chapter 3, 3 reliability research of DTMNs will be discussed. In chapter 4, we will discuss algorithm, which delivers location information only in order to select a path with higher delivery rate when transmitting to detected node, and its performance. Finally, in chapter 5, we will discuss conclusion and further research.

2 Relevant Research

DTMNs(Delay Tolerant Mobile Networks) is a specific category of DTNs. Every node is mobile and end-to-end path does not exist between any given two nodes. Every node in DTMNs in blind and autonomous.

2.1 Delay Tolerant Networks

MANETs introduced various protocols in order to find end-to-end path between nodes, however they mainly focused on routing[1-3]. In extreme environments, a situation where end-to-end path does not exist, is very common. They depend on store-on-forward type for message transmission in divided mobile network and scattered sensor network to research and propose a solution.

These transmission technique may generate numerous message transmission technology under a couple of assumption. For example, assuming that the entire node movement can be controled, using "data mules" to collect data from static sensors is one of them[9]. Another is using "ferries" to find optimal path for message transmission between scattered static nodes[10]. For last, Epidemic Routing technique propose a method, which simply spread messages for message transmission in networks[11]. These technique all focused on message transmission and routing technology in extreme environments.

Delay Tolerant Networking Research Group (DTNRG) introduced DTNs structure for connections between heterogeneous network connection in extreme environments[12]. DTNs provides overlay structure for interoperability between heterogeneous network in extreme environments. Bundle Layer protocol is introduced to deal with new network adventures using store-and-forward research as in past various research[13-15]. "custodian" system is also proposed based on custody transfer concept[16].

2.2 The Concept of DTMNs

DTMNs is a specific type of DTNs, which assumes that every nodes in the network move and end-to-end path does not exist between two nodes in network[16]. Because nodes are scattered and its mobility under this environments, we will see each node as a "region" in regards with typical DTNs structure[4]. Similarly, each node performs overlay bundle action in order to transmit message as DTN gateway. In this paper,

two assumptions are made with nodes in DTMNs. First, nodes exist without plan or information. In other words, nodes in the network does not have any information regarding condition, location or movement pattern about other nodes. Secondly, nodes are autonomous. Again, each nodes perform self-dependent control[8][17]. Under two assumption, Flooding technique is mostly used to exchange and transmit messages. For example, there is Epidemic Routing technique, which simply spread for message transmission in the network[11].

3 Reliability Research

In this chapter, we will introduce 3 reliability researches. Firstly, we will introduce hop-by-hop DTMNs reliability research. Then, we will discuss the other research for end-to-end message transmission in DTMNs.

3.1 Hop-by-Hop

Hop-by-Hop reliability is first introduced in the typical DTNs[4]. In this concept, a message is transmitted via region on the path to the destination. Each region is represented as hop. Gateways located at the boundary in this region works as container and are responsible for trusted message transmission via region[5]. Thus, there isn't any end-to-end acknowledge message in this theory, source node only requires to know whether the next gateway received a message and assumes the gateway would handle the rest of the process[5, 8].

<Figure 1> demonstrates hop-by-hop reliability model process in DTMNs S is source node, F is middle node and U is the final destination node. An arbitrary node receives a message successfully, it sends acknowledge message. An ACK message is returned as an acknowledge to the message receive. Source node as well as middle node voluntarily spread message to numbers of nodes. Given that there are enough time and mobility, source node assumes that a message would arrive to the destination node successfully. Although hop-by-hop lacks in end-to-end reliability, it minimizes the time which a message stays in the buffer of source node. Because source node does not require to wait for end-to-end ACK message.[5, 8]

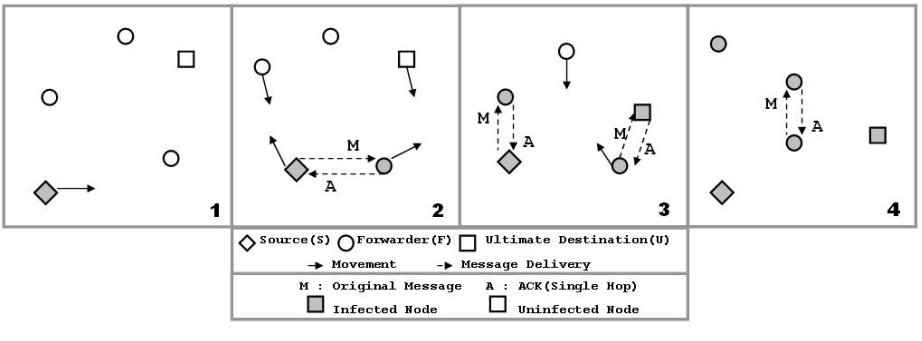


Fig. 1. hop-by-hop reliability model process in DTMNs [8]

3.2 Active Receipt

Problem in Hop-by-Hop reliability lies in lack of end-to-end reliability. It does not guarantee end-to-end reliability in case of node destruction or nature disaster. In order to overcome the disadvantage of the hop-by-hop reliability, we introduce Active Receipt[5, 8].

Active Receipt is based on end-to-end ACK message. Destination node receives a message from source node and generates "receipt" and sends "receipt" to the source node "actively". The term “actively” means nodes deal with new message, receipt, for message transmission.

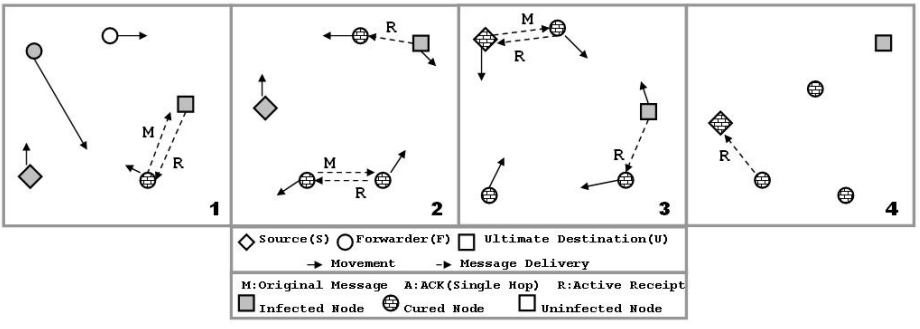


Fig. 2. Active Receipt reliability model in DTMNs [8]

<Figure 2> describes Active Receipt working model in DTMNs. Assuming that the final destination node have received a message under DTMNs message transmission process, it shows the process of delivering ACK message of the message transmitted to the source node. After the destination node receives a message, it spreads "Active Receipt". What can be observed in this process is infected nodes with "Active Receipt" message stop transmitting messages in the network and prevent itself from infected again with messages. Although message infection stops through the network, R, itself continues to be transferred until the timeout or TTL. Traditional hop-by-hop technique infects neighboring nodes with message unconditionally for given timeout, in "Active Receipt" reliability working model reduces unnecessary message transmission once the message is transmitted to the destination node. Therefore, "Active Receipt" transmission costs less than message transmission in hop-by-hop reliability working model. [5, 8]

3.3 Passive Receipt

Active Receipt has a disadvantage, which costs too much resources in order to guarantee end-to-end reliability. In order to transmit a message to one destination, infected nodes in the network require to deliver two message. In order to outcome resource consuming issue, we introduce passive receipt[5, 18].

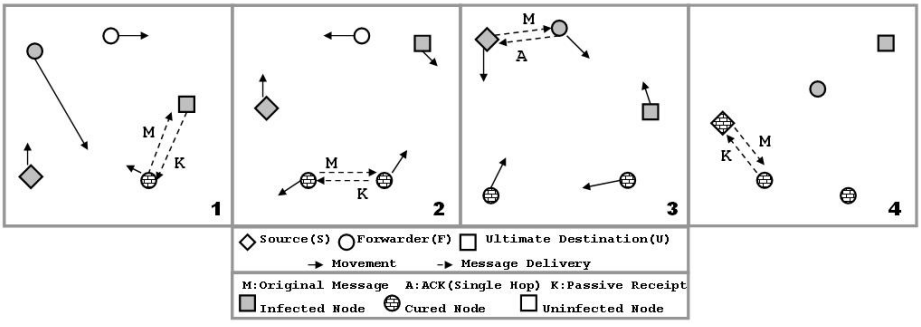


Fig. 3. Passive Receipt reliability model in DTMNs [8]

<Figure 3> describes Passive Receipt working model distinctively. It resembles with <Figure 2> but instead of generating normal "Active Receipt", it sends Kill message(K) to the infected node in order to make them stop message transmission. This working process reduces unnecessary message transmission in hop-by-hop reliability working model and "Active Receipt", also reduces costs significantly.[5,8]

4 Proposal of Reliability Algorithm Using Location Information and Performance Analysis

Hop-by-Hop reliability model only presents basic working model and it does not guarantee the reliability if the end-to-end path does not exist. It also sends out a confirm message in the process of data transmission, it generates unnecessary traffics. In Active Receipt model, there is a risk of transmitted data buffer overflow in order to guarantee end-to-end reliability, it costs too much resources as it spreads one or two messages for reliability guarantee.

4.1 Message Definition and Structure

Let us explain types of message before explaining proposed technique in this paper. This technique employs 3 types of messages as suggested in <Table 1>. The structure of each message type is following.

Table 1. Definition of message

Type	Description
Message	Data message
Passive Receipt	Message for abort transmission to infected node
Information Table	Message includes location informaton

Message is a message for performing data transmission from sender, it includes serial number from sender to decide message loss.

Passive Receipt is a message to ask its infected neighbor nodes to stop message infection, once the final destination nodes receives location information table.

Information Table is a table that includes location information to set a path between source node and destination node. The source node infects its neighbor nodes with its current location and the address of destination node and each infected node would infect its neighbor node with previously received information and its location information. Eventually, when it arrives at the final destination node, the location information table would be completed. Each nodes are moving continuously, the table also contains information about moving path and speed of each nodes.

4.2 Message Transmission Using Implicit ACK

In Hop-by-Hop data transmission, which demonstrates basic working model, the middle nodes requires storage memory for data transmission and confirmation using ACK messages. These ACK messages generated between the source node and the destination node would generate a lot of unnecessary traffics and brings additional packet overhead to the network.

In this paper, we propose a method, which reduces control loads by reducing ACK message using implicit ACK message. An infected node from the sender node infects its neighbor nodes, sender node is also affected indirectly, this will be considered as an ACK message. We will call it an implicit ACK message. In wireless communication, when receiver node forward or relay a message to the next node, the sender node also receives this message indirectly. In multicast communication under symmetry signal transfer environments, this would happen naturally without additional work[18]. We will apply this implicit ACK method as a signal for ACK to signal sender nodes that it has received forwarded packet from sender successfully.

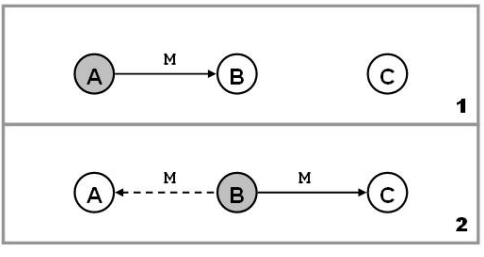


Fig. 4. Implicit ACK working model

<Figure 4> describes the working process of implicit ACK. Sender A transfers a message to Receiver B. B spreads a message around in order to infect its neighboring nodes. At this moment, the sender node A receives a forwarded packet from Receiver B automatically and indirectly and by using this information, it is possible to confirm a packet is transferred successfully from A to B. Unlike Hop-by-Hop reliability transfer, Receiver node B does not required to send back ACK message, resulting in reducing unnecessary traffics.

4.3 Epidemic Based Location Information Table Distribution and Path Setting

The proposed algorithm in paper, in order to improve the transmission reliability, the source node sends location information only to detected nodes and make it choose a path with the highest transmit possibility.

The algorithm described above can be defined as following

```
1. Type SourceNode =
2.     record
3.         NodeId, NodeType, Speed, Path, Hop_Count;
4.         DistanceVector : Information Table
5.     end
6.
7. Global var
8.     InfomationTable : Information Table of   Neighbor Node for Path_Create
9.
10.Procedure Table_Construction
11.    do
12.        if N received from node S
13.            Add S's distance to N's DistanceVector;
14.            Send N's Neighbor Node;
15.        fi;
16.    od;
17.
18.Procedure Path_Create and Message_Send
19.    do
20.        if D received from node N
21.            Add N's distance to D's DistanceVector;
22.            Send D's InformationTable Path to Source;
23.         fi;
24.        if N received from node N
25.            Stop Forwarding InforamtinTable
26.        fi;
27.        if S received from N
28.            Send Message to Destination Node;
29.        fi;
30.    od;
```

The source node and infected nodes in the process does not infect its neighbor nodes with a message but infect them with its location information. The location information may contain the location of the source node, moving path, speed. Infected node adds its location information and transfers it to the next node. Eventually, when destination node receives the location information, it would follow the information in the location information table to deliver location information table to the source node as an ACK

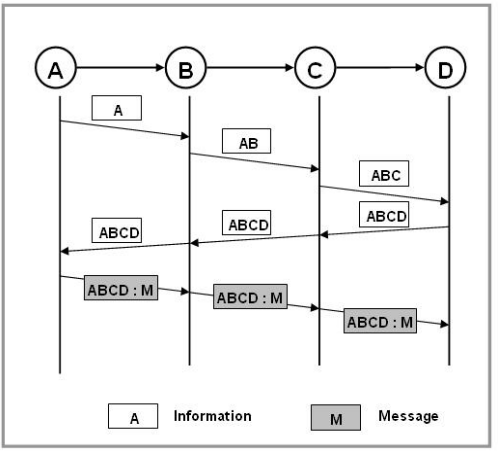


Fig. 5. The process of message generation and transmission of each nodes in the proposed reliability algorithm

message. Then the source node may use the location information table to transfer a message to the destination node.

<Figure 5> demonstrates the process of generating location information table and transmission for each nodes in order to deliver a message and set a path in the proposed reliability algorithm.

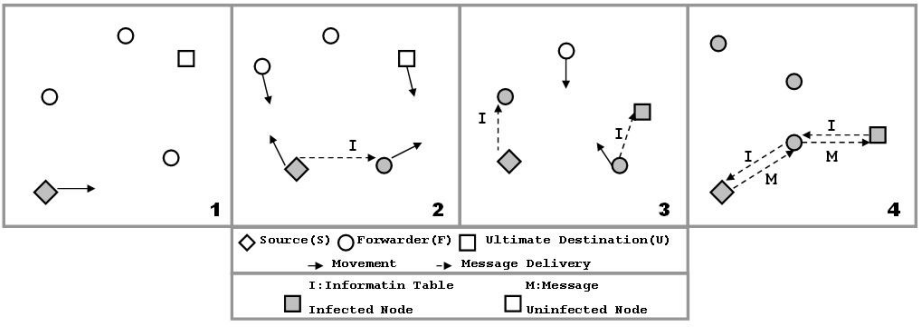


Fig. 6. Location information message transmission and path setting process in DTMNs

<Figure 6> shows reliability algorithm model using the process of location information message transmission and path setting in DTMNs. The reliability algorithm does not initiate the process with the whole message, it only spreads the location information table to reduce the size of storage for each nodes. The source node also infects its neighbor node only with location information table, infected nodes only adds its location information to the table and infects its neighbors. Therefore, each node does not require to send ACK message every time, it would prevent unnecessary traffics from generating. The source node and the destination

node delivers message via a pre-set path, thus preventing unnecessary re-transmission or multiple path selection.

4.4 Performance Evaluation

This test is performed using DTNs simulator, The Opportunistic network Environment Simulator(The ONE), developed by Finland Helsinki University[19]. <Table 2> shows the environmental variables for reliability algorithm test using location information shown in Figure 4 ~ 6.

Table 2. Simulation parameters

Variable	Value
Topology size	10 km^2
Node number	10, 25, 50
Simulation period	1hour, 3hours, 6hours
Beacon interval	1 sec
Retransmission wait time	10 secs

<Figure 7> and <Figure 8> compares the number of message exchange by changing the number of nodes in the test

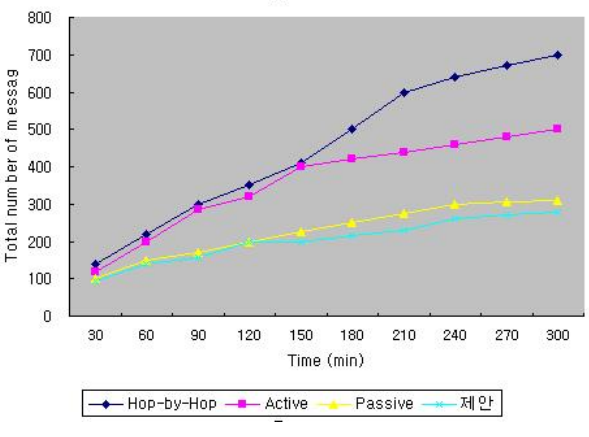


Fig. 7. 25 nodes

As shown in <Figure 7>, the number of message exchange reduced when using proposed algorithm. However, the difference is reduces after a certain amount of time, this converged into similar result as the test environments are randomized.

<Figure 8> also confirms that the number of message exchange between nodes is reduced.

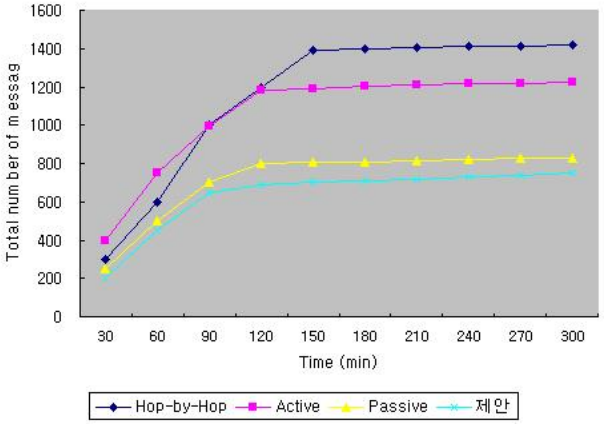


Fig. 8. 50 nodes

5 Conclusion and Further Research

Explosive increase of wireless device and user's mobility requires us to take another technical adventure in other network environments.

New network environments such as satellite network, earth-to-planet communication, military network, isolated remote village and disaster rescue are more prominent with the recent nature disaster events and it is required to communicate in extreme network environments. In order to solve mobility issue in extreme network environments, MANETs proposed various technique focusing on routing and DTNRG proposed new DTNs structure.

In this paper, we have considered main aspects of transfer layer in DTMNs, a specific type of DTNs, which assumes all nodes move in the network and there is no definitive end-to-end path between two nodes. We also introduced 3 reliability algorithm, which are hop-by-hop, active receipt and passive receipt reliability algorithm. We were also able to choose the best reliability by using DTMNs complexity. For example, hop-by-hop reliability algorithm is too simple and the choice between active-receipt and passive-receipt can be made by cost and delay with the order of priority.

The proposed algorithm in this paper infects neighbor nodes with location information table to decide a path, each node does not require to store unnecessary message. The final destination node may refer to the complete location information table and sends the source node in the form of ACK message, each node does not require to spread ACK message as well. It also decide the path with the highest transmission possibility, reliability is greatly increased than traditional algorithm.

We may consider to research on reliability multimedia data transfer in wired network using DTNs bundle protocol and store-and-forward method, we may also continue various research, which considers cost and delay issue in the sensor network while guaranteeing the reliability of sensors.

References

1. Perkins, C.: Ad-hoc On-Demand Distance Vector Routing. In: IEEE Workshop on Mobile Computing Systems and Applications, New Orleans, LA, pp. 90–100 (February 1999)
2. Perkins, C., Bhagwat, P.: Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers. In: ACM SIGCOMM, London, England, pp. 234–244 (October 1994)
3. Royer, E., Toh, C.: A Review of Current Routing Protocols for Ad-hoc Mobile Wireless Networks. IEEE Personal Communications Magazine 6(2), 46–55 (1999)
4. Fall, K.: A Delay-Tolerant Network Architecture for Challenged Internets. In: ACM SIGCOMM, Karlsruhe, Germany (August 2003)
5. Fall, K., Hong, W., Madden, S.: Custody Transfer for Reliable Delivery in Delay Tolerant Networks. Intel Research, Berkeley-TR-03-030 (July 2003)
6. Jain, S., Fall, K., Patra, R.: Routing in a Delay Tolerant Network. In: ACM SIGCOMM, Portland, OR (August 2004)
7. Cerf, V., et al.: Delay-Tolerant Network Architecture. IETF RFC 4838, Informational (April 2007)
8. Harras, K., Almeroth, K.: Transport Layer Issues in Delay Tolerant Mobile Networks. In: IFIP Networking Conference, Coimbra, Portugal (May 2006)
9. Shah, R., Roy, S., Jain, S., Brunette, W.: Data MULEs: Modeling a Three-Tier Architecture for Sparse Sensor Networks. In: IEEE International Workshop on Sensor Network Protocols and Applications, Anchorage, AK (2003)
10. Zhao, W., Ammar, M., Zegura, E.: Controlling the Mobility of Multiple Data Transport Ferries in a Delay-Tolerant Network. In: IEEE INFOCOM, Miami, FL (March 2005)
11. Vahdat, A., Becker, D.: Epidemic Routing for Partially Connected Ad Hoc Networks. Technical Report CS-200006, Duke University (April 2000)
12. DTNRG, A Delay-Tolerant Networking Research Group, `http://www.dtnrg.org/`
13. Scott, K., Burleigh, S.: Bundle Protocol Specification. IETF RFC 5050, Experimental (November 2007)
14. Farrell, S., Cahill, V., Geraghty, D., Humphreys, I., McDonald, P.: When TCP Breaks: Delay and Disruption-Tolerant Networking. IEEE Internet Computing 10(4), 72–78 (2006)
15. Warthman, F.: Delay-Tolerant Networks (DTNs): A Tutorial V1.1 (Mar 2003)
16. Harras, K., Almeroth, K., Belding-Royer, E.: Delay Tolerant Mobile Networks(DTMNs): Controlled Flooding Schemes in Sparse Mobile Networks. In: IFIP Networking, Waterloo, Canada (May 2005)
17. Harras, K.A., Almeroth, K.C.: Inter-Regional Messenger Scheduling in Delay Tolerant Mobile Networks. In: Proceedings of the 2006 International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM 2006), pp. 93–102 (June 2006)
18. Sung, H.K., Yang, H., Chang, Y.P.: Reliable Multicasting with Implicit ACK and Indirect Recovery in Wireless Sensor Networks. Journal of Information Science, Telecommunications 35(2), 215–226 (2008)
19. The ONE, The Opportunistic Network Enviornment simulator, `http://www.netlab.tkk.fi/tutkimus/dtn/theone`

A Study on Automatic Analysis of Social Network Services Using Opinion Mining*

Ye Jin Kwon and Young Bom Park

Information Architecture Laboratory, Department of Computer Science and Engineering,
Dankook University, Cheonan, Korea
kwon6838@naver.com, ybpark@dankook.ac.kr

Abstract. By the arrival of the social network services, the online community changed from information-centered to relation-centered. so the research which analyses relational-centered online community is being actively in progress. However, since the analysing methods of the social network services designed to handle massive data, are not formalized yet, only the partial analysis results is being used to extract the conclusions. In this study, the massive data generated by the social network services are automatically classified using SVM, and the methodology which analyses relational tendency of each user's interaction using opinion mining is proposed.

Keywords: SNA(Social Network Analysis), opinion mining, social network service, SVM(Support Vector Machine).

1 Introduction

Due to the emergence of Social Network Service, the wide range of online communication is possible based on a wide variety of relationship. On one hand, the existing online community based on internet searches uses network services only for the acceptance and understanding of information. On the other hand, the community-based social networking services changes the online community into person-to-person relation-centered community. Namely, Users of social network service show build relationships and participate actively in the communication. The social influence of social network services is larger than a personal blog or mini homepage.[1]

This phenomenon results fast and active research to analysis dynamic relationship in social network services. There several related researches are made in this field such as, "SNS social involvement and as a result is formed of network connection property and unite property"(Shahm McLeod, & Yoon 2001; Boyd &Ellison, 2008), "The modeling study of the structure and characteristics about SNS social networking" (Monger & Contractor, 2003; Trier, 2008), and "SNS research in measurement techniques of influence" (Meeyyoung Cha & hamed Haddadi 2011).

However, studies so far has not been formalized to handle the vast amounts of data analysis methods. Analysts are adhered the classic approach way to review the entire

* This research was supported by National Research Foundation of Korea(KRF-2009-322-A00106).

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 240–248, 2011.

data and derive conclusions. For this reason, actual extensive form of relationships of social network service is not analyzed correctly. the analysis with only a fraction of the data has been studied. Thus, in this study, the automated social network service analysis method to classify whether positive or negative of Interaction between users is proposed for automatic classifying tool development. And measures on the processing of large amounts of data are provide.

2 Related Work

2.1 Interaction

A general sense of interaction exchange act, namely people or things affect each other send and receive act says[2]. What is over online interaction like social network service process mainly in short sentences that communication between users. This type of interaction, the relationship between the user can be expected and ifer. If an active exchange with the other, familiarity with the similarities between the two can be speculated.

According to a recent study, Bearing in the real world to measure the strength of those relationships, such factors as the similarity or familiarity can be applied in online like social network service turned out. In that study the relationship strength of between people's through 11 kinds of indicators like frequency of contact, familiarity, enthusiasm for the relationship, advice, desire of being colleagues, a variety of topics, the relationship duration, dependencies, emotional strength, reliability, good feeling socializing were measured[3].

Accordingly, in this study the interactions within social network service in a variety of measure and calculate ways was designed the process. Through It a simple data model was visualization.

2.2 Opinion Mining

Opinion mining means that users create documents on the Web by analyzing the subjective opinions of the author and that derive the operations. Which to base technology is such as natural language processing, text mining, statistical areas[4].

Opinion mining is divided in three phases. First what it believes is best to define the various features and that derive steps 'feature extraction step', the second What do you mean by the information sources was used in the extracted features and vocabulary that represents the opinion judgments and classified steps to 'opinion classification step', lastly revealed opinions tendency opinions of information summary and whole contents of the information effectively delivered to the user the steps to 'summary and represent steps' be divided[4].

Namely, the opinion mining is determines the direction of whether positive or negative attitude that words, sentences, and documents. This study, design suggest that produces the results determining the direction on the attitude of the interaction among the users.

2.3 SVM (Support Vector Machine)

SVM is developed the learning technique by staticiansVapnik in 1998. Nonlinear problems associated with the input space matches the high-dimensional feature space

the linear problem. So to analyze mathematically easy[5]. Also, do not have a lot of parameters need to adjust the number of relatively simple can be identify factors that affect learning.

And by minimizing the structural risk can be Away from overrated relevant problem. Because the learning process to minimize a convex function can obtain global optimal solution point has attracted attention that the performance is good compared with artificial neural networks machine learning techniques.

Accordingly, in this study the interactions within social network service in a variety of measure and calculate ways was designed the process. Through It a simple data model was visualization.

The principle of SVM can be explained by two major.

First, The principle is' maximum margin classification'. SVM basically the classifier called support vector that most distant from points on the boundary of the two groups is designed to look for. In this way, linearly constrained quadratic Plan (QP) problem (linearly constrained quadratic programming) that best can be divided into two groups by formulation is designed to be found.

The second core operating principle of SVM is 'mapping data into higher dimensional space'. Linear classifier problems that are not classified as In low-dimensional space if the space to mapping higher dimensions linear classifier can be classified better as it works, another key principle of SVM [7].

3 System Architecture Is Based on the Interaction

3.1 Feature Data Selection

This study, each interaction type analysis of based on the tweeter, and accordingly by setting the metrics for the design of data analysis was performed.

First, categorizes the type of interaction of Twitter as the next Table 1.is shown.

Table 1. Android API extract list

Type	Outline
tweet	Means that the 140-character Twitter posting. To Twitter was created within all of a short 140 tweet is called.
Reply	The tweet that in front of other people ID to paste a '@' start with @ID. For any post that is created when you want to tell my opinion.
Mention	In the middle of posts in the form of @ ID mentioned in that the user tweets. If you want to mention by specifying a particular person in the message pass is the ability to use.
Direct Message (DM)	Tweet that sender and receiver can see only two in private.someone who personal story is used to send and receive.
Retweet (RT)	my person who following person's tweet of pass back to my followers. If you have information that is valuable and worth inform my followers are used when you want to share.

The basic types of interactions in Twitter as above can be classified into five. The default type of interaction is just one of the five that appear Accordingly, the relationship between users and the very diverse and complex appear.

In this study, a variety of users in relation to the interaction between the pay measure and each of the continuous interaction among the users led to the prediction when the similarity between two users, and produce the results to derive.

Social network service, the exchange and flow of information passing between from Twitter can be considered passed between users tweet.

Thus for in order to measure the density of among users such as each tweets (tweet) delivered between users (retweet), mention, reply the type of interaction should be based. These interactions between users of Twitter represents an overview of the Fig.1.looks like.

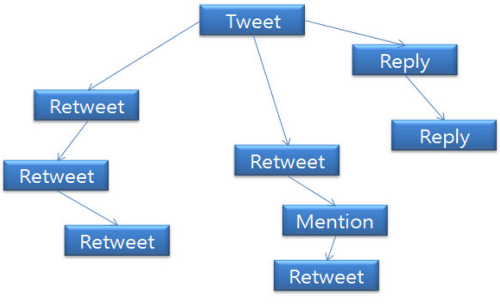


Fig. 1. Interactions between users of Twitter

Above Fig1 look at the, about A specific tweets (Tweet) the user's reaction and subsequent spread thus any the real interaction between each user, unknown is been made. Based on this information, the actual extent of the interaction between the user has indicated that the Fig.2.is shown.

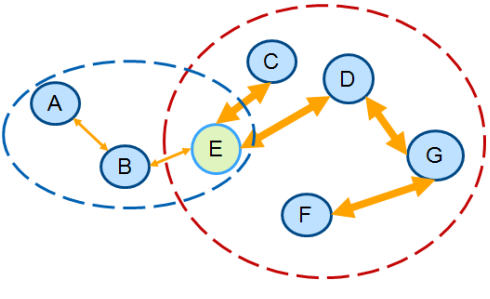


Fig. 2. Interactions between users of Twitter

In the picture above, depending on the degree of each interaction was expressed differently than the thickness of the arrows. Actual users 'E' and the many interactions that make up the group and does not form a lot of interaction is to separate the groups.

Accordingly, Users 'E' to a lot of interaction with the user group prediction will be that tend to be similar to user 'E'.

So far, a specific teewt as an indicator of the interaction representation of the relationship Table 2 is shown.

Table 2. Selecting of Feature Data

Group	Feature	Explain
Response	reply-tweet	check of the reply
	reply-time	Written time of replies
	mention-tweet	check of the mention tweet
	mention-time	Written time of mention tweet
	retweet-others	check of the retweet
	retweet-time	Written time of retweet
Relation	relation-distance	distance of relation
	direct-relation	direct-relation
	indirect-relation	indirect-relation
Interation	response-reply	check of the response
	reponse-time	Written time of response
	interaction count	count of the interaction

First, select the first of specific tweet then that tweet some was the reaction and the reaction time is how much is formatted. The detailed data format divided into reply in the form and mention, retweet. For each reaction of the actual time response and the number of availability is formatted.

The second response to the specific tweets when listening for the degree of relationship is represented by the data. Whether it is directly related to, posted the specific tweet directly with the user followers - follwing relation, or by other users tweets that you have received is to measure passes.

Third, if the actual reaction of tweets, they be made for this reaction whether has the response or not is formatted as data. For the specific tweets, such as responses being made for the reaction have been made in response to and response time for the reaction how much information is represented.

4 System Configuration

A system of that based on data from the selected feature data processing and to predict the similarity between users looking at the overall configuration, and then Fig. 3. looks like.

The first, that you want to analyze determined data is converted to a feature data type. Then for the prediction similarity creates a data model. Algorithms that form the data model representative prediction algorithm SVM-based is generated.

Opinion mining also the interaction between the user tendency to identify data dictionary to include in the process. Based on the generated data model that is generated as a result of social networking services has a large amount of data classification to derive the data produce results.

Prediction algorithms that derive a large amount of data processed by the SVM can be represented.

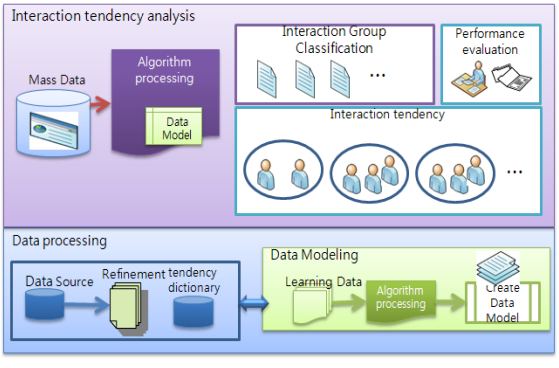


Fig. 3. System Architecture

Classified data is generated based on the results in the social network services of interaction among the users. That tendency of the interaction between users of data is used as an inference.

Then all courses of that formed based on the interaction of each tweet interaction tendency compared with pre-stored data dictionary and the tendency deduce between each users through large amounts of data could be analyzed automatically processed. In addition, the processing results to some extent that performance can be expressed add a module that can be evaluated. So predicted reliability of the analysis results be ensure.

5 Experiment and Results

5.1 The Experimental Data Classification

In this study, test data set of tweeter simply made the feature data of based on interaction. Then classification using SVM Light which experiments were carried out. Selected feature data is numerically represented of interactions within social network service.

```
1 1 1:1 2:0.12 3:1 4:0.74 5:1 6:1.23 7:2 8:1 9:0 10:3 11:0.77 12:6
2 1 1:1 2:0.09 3:1 4:0.45 5:1 6:0.69 7:3 8:1 9:0 10:3 11:0.11 12:4
3 1 1:1 2:0.23 3:1 4:1.34 5:1 6:0.48 7:1 8:1 9:0 10:5 11:0.35 12:3
4 1 1:1 2:0.11 3:0 4:1.47 5:1 6:0.25 7:3 8:1 9:0 10:5 11:0.52 12:5
5 1 1:1 2:0.53 3:0 4:0.1 5:1 6:0.91 7:3 8:1 9:0 10:4 11:0.49 12:5
6 1 1:1 2:0.12 3:1 4:0.9 5:1 6:0.1 7:2 8:1 9:0 10:4 11:0.96 12:5
7 1 1:1 2:0.74 3:0 4:1.41 5:0 6:0.82 7:2 8:1 9:0 10:1 11:0.84 12:8
8 1 1:1 2:0.23 3:1 4:0.38 5:1 6:0.8 7:1 8:0 9:1 10:3 11:0.57 12:8
9 1 1:1 2:0.66 3:1 4:1.08 5:1 6:0.83 7:2 8:1 9:0 10:3 11:1.3 12:5
10 1 1:1 2:0.35 3:1 4:0.32 5:1 6:0.15 7:2 8:1 9:0 10:5 11:0.91 12:7
                    ⋮
```

Fig. 4. SVM Data Set

Next Fig.4. purified to make the classification indicates Data Set. Above, refined data set of is made for users of 143 people the tweeter. As defined in Table 2., this represents of among users of the reaction, the degree of relationship, the degree of interaction.

Using Dataset with SVM Light when you create a data model, next Fig5.and the data model is defined. The generated data is used to based on the interaction of the user to automatically analyze. In addition, the analysis of a large number of users for processing in less time is the basic definition.

```
1  SVM-light Version V6.01
2  0 # kernel type
3  3 # kernel parameter -d
4  1 # kernel parameter -g
5  1 # kernel parameter -s
6  1 # kernel parameter -r
7  empty# kernel parameter -u
8  12 # highest feature index
9  143 # number of training documents
10 2 # number of support vectors plus 1
11 -0.99999767 # threshold b, each following
12 line is a SV (starting with alpha*y)
13 4.1825499832909933e-008 1:1 2:0.12 3:1
14 4:0.74000001 5:1 6:1.23 7:2 8:1 9:0 10:3
15 11:0.76999998 12:6 #
16
```

Fig. 5. Define of the Data model

Based on The generated data would you have the data to predict the classification, next like <Figure 6> the result is displayed.

As can be seen from the above results, through the test data is classified automatically to active interaction user group and non-active interaction user group. To know is the result that analysis of a specific user with a relational-oriented. And

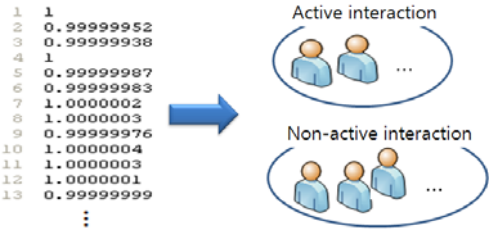


Fig. 6. Predicted the final data

this is know that relation analysis of the interaction between the user and can be done automatically

5.2 The Trend Predictions of the Using Opinion Mining

Section 5.1 of the entire system procedures, based on the interaction among the users was into two group classification. As a result, happens group of active interaction can extract. Based on the results obtained for the group, if you want using opinion mining is predictions of the tendency next Fig.7.and will follow the same procedure.

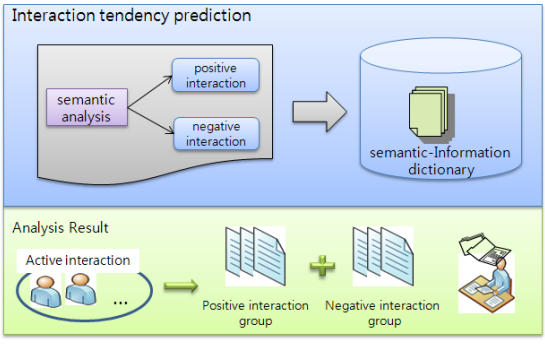


Fig. 7. Opinion Mining Prediction Procedure

First, the meaning of data that positive interaction and negative interactions collect and create a data dictionary. Through written data dictionary, analysis When a positive means to interact and a negative means to interact is possible. Accordingly, group on the extracted in Section 5.1 that interaction is actively happens divide of semantically.

6 Conclusion

Until now, Up to now, we formalized large amounts of data that generated as a result of social network service, and automated tendency assessment analysis method was designed base on that data. user interaction classification possibility test was

performed through formalized test data. Finally, through the grouped data, the user interaction tendency prediction process using opinion mining was proposed.

The fast data processing for social network service user tendency is expected, when proposed automated social network service analysis method using opinion mining is applied. Furthermore, Since the user tendency prediction is performed based on the feature data which is generated by the interactions between users and the characteristics of social network services, the proposed method shows the possibility of faster user relation analysis method in relational-centered online community.

As a future study, the automated social network service analysis tool will be constructed based on proposed method. And the testing and evaluation will be performed using this tools. Our ext For of Satisfaction and reliability will conduct the study of ensuring the satisfiability and reliability of the analyzed data will be addressed.

References

1. Wei, L.: Filter blog cs. Personal Journals: Understanding the Knowledge Production Gap on the Internet. Journal of Computer-Mediated Communication 14, 532–558 (2009)
2. Oxford English Dictionary Online Version, 2nd edn.
3. Nepusz, T., Bazsoo, F.: Measuring tie-strength in virtual social networks (2007)
4. Bo, P., Lillian, L.: Opinion mining and Sentiment Analysis
5. Hearst, M.A., Dumais, S.T., Osman, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intelligent System 13(4), 18–28 (1998)
6. Vapnik, V.: Statistical Learning Theory. Wiley (1998)

On the Security of a Robust Watermarking Scheme Based on RDWT-SVD

Huo-Chong Ling[1], Raphael C.-W. Phan[2], and Swee-Huay Heng[1]

[1] Research Group of Cryptography and Information Security,
Centre for Multimedia Security and Signal Processing,
Multimedia University, Malaysia
{hcling,shheng}@mmu.edu.my
[2] Loughborough University, LE11 3TU, United Kingdom
r.phan@lboro.ac.uk

Abstract. Image watermarking schemes allow a cover image to be embedded with a watermark for diverse applications, such as copyright protection and covert communication. Recently, a hybrid image watermarking scheme based on redundant discrete wavelet transform and singular value decomposition is proposed by Lagzian et al. This scheme demonstrates high robustness against common image processing attacks such as geometric attack, JPEG compression, Gaussian noise and histogram equalization. In this paper, we present a fundamental flaw in the scheme that leads to false-positive detection problem and hence, undermines the robustness and suitability of the scheme to be used as a proof of ownership application.

Keywords: watermarking, redundant discrete wavelet transform, singular value decomposition, false-positive detection, robustness, proof of ownership.

1 Introduction

A watermarking scheme [1–8] is one where it is desired to protect the copyright of a content owner by embedding the owner's watermark into the content. In order to prove the ownership of the watermarked content, the owner takes the case of ownership claim to the authority, and proves ownership by performing the watermark detection process on the claimed content to extract his watermark. Therefore, robustness of the watermarking scheme is an important factor, i.e. it should be infeasible for an attacker to remove, modify or prevent the extraction of an embedded watermark without visible distortions of the image.

In this paper, we concentrate on singular value decomposition(SVD)-based watermarking schemes used for proof of ownership application. SVD is a linear algebra scheme that can be used for many applications, particularly in image compression [9]. Suppose an N-by-N image matrix I with rank $r \leq N$. The SVD of I is defined as $I = USV^T = \sum_{i=1}^{r} u_i s_i v_i^T$ where S is an N-by-N diagonal matrix containing singular values s_i satisfying $s_1 \geq s_2 \geq \ldots \geq s_r > s_{r+1} = \ldots = s_N = 0$,

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 249–256, 2011.

and U and V are N-by-N orthogonal vectors. V^T is the adjoint (transpose and conjugate) of the N-by-N matrix V. Since the SVs are arranged in decreasing order, the last terms will have the least affect on the overall image.

Several SVD-based watermarking schemes have been proposed [1–8] in the past years. The most popularly cited scheme is by Liu and Tan [7] that proposed to insert the watermark into the SVD domain of the cover image, and demonstrated its high robustness against image distortion. However, Zhang and Li [10] and Rykaczewski [11] proved that the Liu-Tan scheme is fundamentally flaw in its algorithm design and thus, is not suitable to be used for proof of ownership application.

In this paper, we furthermore analyze robustness of a hybrid SVD-based watermarking scheme proposed by Lagzian et al. [5] that uses not just SVD but also redundant discrete wavelet transform (RDWT). In Sect. 2, we recall the basics of the scheme proposed by Lagzian et al. We then present the flaws on the scheme in Sect. 3 that invalidate the security claim of the designers. Experimental results verifying our flaws are given in Sect. 4, and Sect. 5 concludes this paper.

2 RDWT-SVD Based Watermarking Scheme

Lagzian et al. [5] proposed a watermarking scheme based on RDWT and SVD, and demonstrated high robustness against common image processing attacks, such as geometric attack, JPEG compression, Gaussian noise and histogram equalization. Lagzian et al. also showed in their experimental results section that their scheme tended to be more robust than the DWT-SVD based watermarking scheme. Another advantage of the scheme is the possibility of embedding a watermark up to the size of the cover image as compared to DWT-SVD scheme. The watermark embedding process of Lagzian et al. scheme is as follows:

E1. Apply RDWT to the cover image I to decompose it into LL, HL, LH and HH subbands.

E2. Apply SVD to the low frequency LL subband of the cover image as:

$$I_L = USV^T. \tag{1}$$

where V^T denotes the transpose of V.

E3. Apply RDWT to the watermark W to decompose it into LL, HL, LH and HH subbands

E4. Apply SVD to the low frequency LL subband of the watermark as:

$$W_L = U_W S_W {V_W}^T. \tag{2}$$

E5. Modify the singular values of the LL subband of the cover image, S with the singular values of the LL subband of the watermark, S_W to obtain the modified singular values S^* as:

$$S^* = S + \alpha S_W. \tag{3}$$

where α is the scaling factor that controls the strength of the embedded watermark.

E6. Apply inverse SVD on the transformed cover image with the modified singular values S^* to obtain the modified low frequency subband ${I_L}^*$ as:

$$ {I_L}^* = U S^* V^T. \tag{4} $$

E7. Apply inverse RDWT using the modified coefficients of the low frequency subband ${I_L}^*$ to obtain the watermarked image I_W.

Take note that the owner needs to keep U_W, V_W and S from this embedding process, to be used subsequently during the watermark extraction process so that the extractor can properly extract his watermark. The watermark extraction process is as follows:

X1. Using RDWT, decompose the possibly distorted watermarked image ${I_W}^*$ into LL, HL, LH and HH subbands.
X2. Apply SVD to low frequency subband LL of ${I_W}^*$ as:

$$ {I_{WL}}^* = U^* S^* {V^*}^T. \tag{5} $$

X3. Extract the singular values of the LL subband of the extracted watermark, ${S_W}^*$ as:

$$ {S_W}^* = (S^* - S)/\alpha. \tag{6} $$

where S is the value obtained from Eq.(1) and is kept by the user.
X4. Apply inverse SVD to ${S_W}^*$, U_W and V_W components to obtain the LL subband of the extracted watermark.

$$ {W_L}^* = U_W {S_W}^* {V_W}^T. \tag{7} $$

where U_W and V_W are provided by the owner.
X5. Apply inverse RDWT using the coefficients of the LL subband, ${W_L}^*$ to obtain the extracted watermark, W^*.

Although Lagzian et al. showed that their scheme is robust against common image processing attacks and achieves better performance than the DWT-SVD based schemes in terms of high peak signal-to-noise ratio (PSNR) and correlation coefficient, their scheme is still vulnerable to false-positive detection problem. In a false-positive detection problem, a watermarked image is proven to contain a particular watermark W_A by an attacker even though W_A has never been embedded in the image in the first place.

In the next section, we show the false-positive detection problem in Lagzian et al. scheme that makes their scheme unsuitable to be used for proof of ownership application, which is important for a watermarking scheme.

2.1 Theoretical Analysis of the Problem

The key point of the false-positive detection problem as mentioned in the previous section is due to a fundamental flaw in its watermark extraction process, in

similar vein to the problem exhibited by Zhang and Li [10] and Rykaczewski [11] on Liu and Tan SVD-based watermarking scheme [7], and Ting [12] on Ganic and Eskicioglu DWT-SVD based watermarking scheme [4].

The flaw lies in Step X4 of the watermark extraction process where the owner has to provide his U_W and V_W components that he obtained from Step E4 of the embedding process, in order to obtain the LL subband of the extracted watermark, ${W_L}^*$. ${W_L}^*$ is then inversed RDWT to obtain the extracted watermark W^*. For an image I, its orthogonal vectors U and V due to SVD contain major information of the image [10, 11]. Therefore, during the watermark extraction process, an attacker can provide his own watermark W_A's orthogonal vector pairs of U_A and V_A to obtain a good estimate of the LL subband of his watermark W_A regardless of what the extracted singular matrix ${S_W}^*$ is in Step X3. The LL subband ${W_{AL}}^*$ is then inversed RDWT to obtain the extracted watermark ${W_A}^*$ that is perceptually similar to his own watermark W_A. As a result, the attacker can equally lay rightful ownership claim on the image because the scheme would extract his watermark.

In other words, the flaw appears because the extraction process of Lagzian et al. scheme makes use of information that is too dependent on the watermark, as shown in Step X4. Therefore, Lagzian et al. scheme is not suitable to be used for proof of ownership application.

2.2 Further Attack

A further attack can be mounted on Lagzian et al. scheme by an attacker to extract his watermark W_A from the watermarked image ${I_W}^*$, even though W_A has never been embedded in ${I_W}^*$ in the first place. The key point to the attack is in Step E5 (i.e. Eq.(3)) of the embedding process, where the modified singular values S^* is replaced by:

$$S^* = S - \alpha S_W. \tag{8}$$

At the end of the embedding process, the attacker would get a fake watermarked image I_{WA}. I_{WA} is then used in the extraction process as follows:

A1. Using RDWT, decompose the possibly distorted watermarked image ${I_W}^*$ into LL, HL, LH and HH subbands.

A2. Apply SVD to low frequency subband LL of ${I_W}^*$ as:

$$ {I_{WL}}^* = U * S * {V^*}^T. \tag{9}$$

A3. Extract the singular values of the LL subband of the extracted watermark, ${S_W}^*$ as:

$$ {S_W}^* = (S^* - S_F)/\alpha_A. \tag{10}$$

where α_A is the attacker's scaling factor and S_F is the singular values obtained from Eq.(11).

$$ I_{FL} = U_F S_F {V_F}^T. \tag{11}$$

I_{FL} is the low frequency LL subband of the fake watermarked image I_{WA} after I_{WA} has gone through the RDWT.

A4. Apply inverse SVD to $S_W{}^*$, U_A and V_A components to obtain the LL subband of the extracted watermark.

$$W_L{}^* = U_A S_W{}^* V_A{}^T. \tag{12}$$

where U_A and V_A are the orthogonal vectors of the attacker's watermark W_A.

A5. Apply inverse RDWT using the coefficients of the LL subband, $W_L{}^*$ to obtain the extracted watermark, $W_A{}^*$.

The extracted watermark, $W_A{}^*$ is perceptually similiar to the attacker's watermark, W_A. In the next section, we show how the flaw and the attack in Sect. 3.1 affect the outcome of the ownership claim via experimental results.

3 Experimental Results

In this section, experiments are carried out to prove that the false-positive detection is feasible in Lagzian et al. scheme [5]. The α value that controls the strength of the embedded watermark is set at 0.05. Fig. 1 shows a gray Lena image of size 200 × 200, an owner's watermark of size 200 × 200 and the watermarked image, respectively. The peak signal-to-noise ratio (PSNR) of the watermarked image in Fig. 1(c) with respect to the cover image in Fig. 1(a) is 30.79 dB and the correlation coefficient (cc) is 0.999. Fig. 2 shows the attacker's watermark and the watermarked image after the attack (PSNR = 25.19 dB, cc = 0.999), respectively. As can be seen from Fig. 2(b) and the PSNR and correlation coefficient values, the modified watermarked image is still perceptually similar to the cover image in Fig. 1(a).

Extraction process is then carried out on the watermarked image in Fig. 2(b) using the owner's orthogonal vectors that are tightly connected to his own watermark in Fig. 1(b), and the attacker's orthogonal vectors that are tightly connected with his watermark in Fig. 2(a). Figs. 3(a) and 3(b) show the results of the extraction process, where a good estimate of the owner's watermark (PSNR = 6.53 dB, cc = 0.995) and the attacker's watermark (PSNR = 38.53 dB, cc = 0.999) can be extracted from Fig. 2(b) successfully. However, the quality of the extracted owner's watermark is far from good with a PSNR of 6.53 dB only.

Attack in Sect. 3.1 is then carried out on the watermarked image in Fig. 1(c). Fig. 4(a) shows the watermarked image at the end of the modified embedding process, as explained in Sect. 3.1. The PSNR and the correlation coefficient obtained are 49.31 dB and 1 respectively. The extraction steps A1 till A5 are performed and Fig. 4(b) shows the final extracted watermark (PSNR = 38.40 dB, cc = 0.999) which is perceptually similar to the attacker's watermark in Fig. 2(a). This attack shows that the attacker's watermark is successfully extracted from the watermarked image in Fig. 1(c) even though it has never been embedded in the image in the first place. Therefore, Lagzian et al. watermarking scheme is not suitable to be used for proof of ownership application.

Fig. 1. (a)Cover image of size 200 × 200 (b)Owner's watermark of size 200 × 200 (c)Watermarked image

Fig. 2. (a)Attacker's watermark of size 200 × 200 (b)Watermarked image after the attack

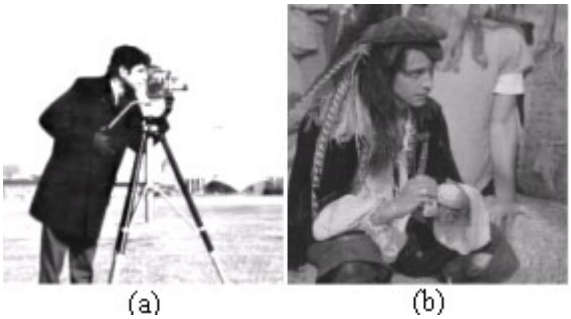


Fig. 3. (a)Extracted watermark using owner's parameters (b)Extracted watermark using attacker's parameters

Fig. 4. (a)Watermarked image after attack in Sect. 3.1 (b)Extracted watermark

4 Conclusions

We have shown that Lagzian et al. scheme [5] has a fundamental flaw that leads to false-positive detection problem, i.e. false detection or extraction of watermark from the watermarked image even if the embedded watermark is different or does not exist at all. The flaw is due to the fact that orthogonal vectors U and V can preserve major information of an image [10, 11] regardless of the singular matrix S and Lagzian et al. scheme makes use of information that is too dependent on the watermark in Step X4. Thus, the attacker takes advantage of this flaw to claim the ownership of the watermarked image even though his watermark does not exist in the claimed image in the first place. This is further supported in the experimental results section. Therefore, Lagzian et al. scheme cannot be used for proof of ownership application.

References

1. Chang, C.C., Tsai, P., Lin, C.C.: SVD-Based Image Watermarking Scheme. Pattern Recognit. Lett. 26, 1577–1586 (2005)
2. Chang, C.-C., Hu, Y.-S., Lin, C.-C.: A Digital Watermarking Scheme Based on Singular Value Decomposition. In: Chen, B., Paterson, M., Zhang, G. (eds.) ESCAPE 2007. LNCS, vol. 4614, pp. 82–93. Springer, Heidelberg (2007)
3. Chang, C.C., Lin, C.C., Hu, Y.S.: An SVD Oriented Watermark Embedding Scheme with High Qualities for the Restored Images. Int. J. Innov. Comput. Inf. Control. 3(2), 609–620 (2007)
4. Ganic, E., Eskicioglu, A.M.: Robust Embedding of Visual Watermarks using Discrete Wavelet Transform and Singular Value Decomposition. J. Electron. Imaging. 14(4), 43004 (2005)
5. Lagzian, S., Soryani, M., Fathy, M.: A New Robust Watermarking Scheme Based on RDWT-SVD. Int. J. Intell. Inf. Process. 2(1), 22–29 (2011)
6. Lai, C.C.: A Digital Watermarking Scheme Based on Singular Value Decomposition and Tiny Genetic Algorithm. Digital Signal Processing 21(4), 522–527 (2011)
7. Liu, R., Tan, T.: An SVD-Based Watermarking Scheme for Protecting Rightful Ownership. IEEE Trans. Multimedia. 4(1), 121–128 (2002)

8. Mohammad, A.A., Alhaj, A., Shaltaf, S.: An Improved SVD-Based Watermarking Scheme for Protecting Rightful Ownership. Signal Processing 88, 2158–2180 (2008)
9. Andrews, H.C., Patterson, C.L.: Singular Value Decomposition(SVD) Image Coding. IEEE Trans. Commun. 24(4), 425–432 (1976)
10. Zhang, X.P., Li, K.: Comments on "An SVD-Based Watermarking Scheme for Protecting Rightful Ownership". IEEE Trans. Multimedia. 7(2), 593–594 (2005)
11. Rykaczewski, R.: Comments on "An SVD-Based Watermarking Scheme for Protecting Rightful Ownership". IEEE Trans. Multimedia. 9(2), 421–423 (2007)
12. Ting, G.C.W.: Ambiguity Attacks on the Ganic-Eskicioglu Robust DWT-SVD Image Watermarking Scheme. In: Won, D.H., Kim, S. (eds.) ICISC 2005. LNCS, vol. 3935, pp. 378–388. Springer, Heidelberg (2006)

Experiment and Verification of Teaching Fractal Geometry Concepts Using a Logo-Based Framework for Elementary School Children

Yeonghae Ko and Namje Park[*,**]

Department of Computer Education, Teachers College, Jeju National University,
61 Iljudong-ro, Jeju-si, Jeju-do, 690-781, Korea
{smakor,namjepark}@jejunu.ac.kr
namjepark@gmail.com

Abstract. Future curriculum of computer education will include the areas of algorithm and programming. As using an educational programming language is an integral part of algorithm and programming education, research should be carried out urgently regarding the use of programming languages. We have to develop and teach with new curriculum which is merging computer subject and the others and we also make curriculum as the children is able to construct knowledge. In this paper, we are going to suggest a new method of teaching computer education to average students wherein elementary school.

Keywords: LOGO, Fractal geometry, Elementary computer education.

1 Introduction

Recently, there is a main stream of education integrating or merging subjects. For example, STEM (Science, Technology, Engineering & Mathmatics) or STEAM (Science, Technology, Engineering, Arts & Mathmatics) education occur in United States and Korean ministry of education, science and technology introduce STEAM in Korean education. STEAM stands for Science, Technology, Engineering, Art and Mathematics. It means integrated education which is merging 5 subject; Science, Technology, Engineering, Art and Mathematics totally for educational purpose. Up to now, computer education in Korea has been operated without any connection with other subjects. In addition, it also has been operated by rote because the class which is taken by an active teacher has been progressed in the way of the direct method of teaching. However, it has many problems from the viewpoint of constructionism. Constructionism is based on two different senses of "construction" of knowledge. First, it is grounded on the idea the children learn by actively constructing new knowledge, not by having information poured into their heads. Second,

[*] This paper is extended from a conference paper presented at the journal of Korean Institute of Information Technology (Vol.9, No.8). The author is deeply grateful to the anonymous reviewers for their valuable suggestions and comments on the first version of this paper.

[**] Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 257–267, 2011.

constructionism asserts that effective learning takes place when the learner is engaged in constructing personally meaningful artifacts. In this respect, we have to developed and teach with new curriculum which is merging computer subject and the others and we also make curriculum as the children is able to construct knowledge. In this paper, we are going to suggest a new method of teaching computer education to average students wherein elementary school.

2 Architecture Theoretical Background

LOGO is an educational computer programming language which is used to teach programming for students. LOGO lies in the distinctness of its feature. First, LOGO gives a chance to make full use of mathematical knowledge and technique and it may facilitate the integration at a cognitive level of the different aspects of the mathematical object. Secondly, LOGO is good for elementary school students because it is based on designing a short and simple language to support the first steps in the learning of programming. Third, LOGO is good for learning programming skills because those are consolidated when students can use the concept that they have developed in previous programming activities in order to deal with problems of increasing complexity.

2.1 Easy to Use

LOGO has so very simple interface that beginner programmer could approach to programming easily. LOGO has simple instruction words and syntax. User can learn programming by starting with five instructions only and working totally with about fifteen instructions that are sufficient for programming any complex behavior of the turtle. Simple syntax and good access to graphics offer good possibilities to study the mistakes one has made and contemplate their reasons and consequences. LOGO is interactive programming language so programmer immediately see the actions of the turtle and can easily revise their instructions if something has gone wrong. In addition, Logo's limited syntactic, technical, and formal requirements and its informative and non-threatening error messages also make this programming language easier to use. These characteristics make Logo an excellent tool for teaching computer science concepts. In this reason, many practice examples show that programming can be very easily learned with Logo not only by adults of any age but also by children from 8 years.

2.2 Motivation

LOGO program show result of instruction to user immediately by using screen. Human tend to depend on sight mainly among the five senses. Besides elementary school students who is target of this article belong to concrete operational period which is one of cognitive development levels in Piaget's theory. They show an extreme interest when they were stimulated by sight. In addition, Logo allows children to create their own miniature world where they learn through experimentation, exploration, and self-directed activity about the cause and effect relationships encountered in computer programming and throughout most of life's experiences. In this reason, LOGO is good for motivation and the more motivated the learner is, the more effectively they will study.

2.3 High Connectivity of Mathematics and LOGO

The programming language LOGO can be used to build bridges between teaching mathematics and computer science. LOGO has several good features for being an excellent tool for studying and teaching mathematics because user can study basic mathematics concept like straight line, angle, shape etc. In addition, Logo is a powerful language that allows for explorations of advanced mathematical topics which is like infinity as well as similarity, congruence and pattern. The programming activities may facilitate the integration at a cognitive level of the different aspects of the mathematical object.

2.4 Using Fractal Geometry in Education

Fractal geometry is not a simple concept but a complex concept in mathematics. It means a shape made up of parts that are the same shape as itself and are of smaller and smaller sizes. It is somewhat difficult to the lower grades of elementary school students but the higher grades can understand and apply it effectively.

3 School Computer Education Using LOGO and Fractals

3.1 Linking Computer Education to Elementary School Mathematics Courses

The ultimate goal of elementary school computer education is using computer education to enhance creativity, problem solving ability, and logical thinking. The same goal applies to mathematics education. A lot of the learning experiences present in computer education can help students achieve the goals of their mathematics education, and vice versa. Therefore, linking elementary school mathematics education to elementary school computer education can produce powerful synergy.

The five sections of elementary mathematics education are numbers and calculations, figures, measurements, probability and statistics, and finally regulation and problem solving ability. Elementary school computer education applied to LOGO and fractals can be connected to several of these sections as shown in table 1.

Table 1. Connection among curriculum of elementary math, LOGO and education of fractal

Grade	Scope and Contents	LOGO Programming
4	· Shape - Different angles and triangles	· Basic Commands - fd :size lt 90 bk :size/2 - fd :size lt 135 bk :size*2/3
4	· Measurement - Angle	
4	· Pattern and Problem solving - making pattern - Pattern and Response	· Fractal geometry (Fig 1), (Fig 5) reference
5	· Shape - congruence - Symmetry	
5	· Pattern and Problem solving - ratio	· Recursive call - logot :size/2 - logoy :size*2/3

First, the basic commands used to compose fractal procedures in LOGO can be related to figures and measurements. Using simple movement commands and changing a turtle's angle allows the angles of simple figures to be understood. Moreover, students can learn how to measure sections by measuring the angles on the screen. There is also a third benefit of using polygons to improve the student's ability to understand figures. Second, recursive calls can be related to the learning of ratios which is related to problem solving. Changing the size of the factor in recursive calls sets the ratio by which to reduce a fractal. The ratio can be confirmed with the eyes because the change in size of the fractal depending on the ratio will be printed on the screen. This learning process allows ratios which are a part of regulation and problem solving ability to be learned. Third, learning fractals can be connected to learning regulations which is a part of problem solving. It is possible to learn about regulations when using procedures to realize fractals because fractals have the features of regulation, repetitiveness, and similarity. Also, part of regulations and problem solving is learning how to make regular patterns and so learning this section is very helpful and useful.

3.2 Development of Education Method for T shaped Fractal Geometrical Theory

The fractal form suggested in figure 1 is designed based on the English alphabet shape. This procedure is designed in a way that each part is drawn by turtle from the cross point in alphabet T. If the recursive call is conducted at each corner of T, a fractal form with a regularity and self similarity is made. If the fractal form is different from the basic form, it would be difficult for an elementary school student to understand. Therefore, the total shape is made in a way that it is similar to the basic form.

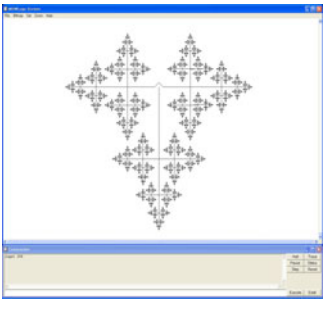


Fig. 1. T fractal geometry screen developed by LOGO

Analysis of newly developed T shaped fractal and algorithm is as follows. The 'Logo-T' procedure is composed of basic commands which an elementary school child may easily understand. The T fractal patterns designed in this study are as follows if they are expressed in Logo language procedure.

```
to logot :size
  if :size<1 [stop]
     bk :size
          logot :size/2
     fd :size
     lt 90
     bk :size/2
          logot :size/2
     fd :size
     rt 180
          logot :size/2
     fd :size/2
     lt 90
end
```

Fig. 2. T fractal procedure screen developed by LOGO Programming

The newly developed T fractal procedure is constructed based on the basic forms. So, the students should understand the basic forms or shapes if they want to learn the procedure. The basic form which is the basic for T fractal is as shown in Figure 3.

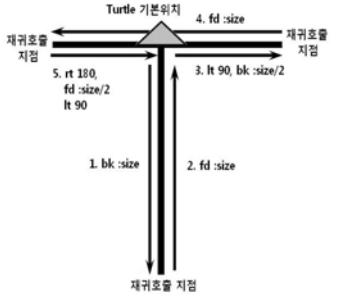


Fig. 3. Fundamental figure of T fractal in LOGO programming

If the procedure is called upon along with the size factor, check first if the size factor is less than 1 using the conditional sentence. The conditional sentence is added so that the infinite loop by recursive call is avoided. The branch at the bottom is to be drawn using the 'bk' command. At the end of it, the recursive call is done to complete the fractal structure. At this time, if the factor value rate is not less than 1 for the recursive call, it falls into the infinite loop. Therefore, the value less than 1 should be used. In addition, if the LOGO turtle is moved to the basic position by using 'fd' command, the branch at the bottom side is completed. Then, using the It command, change the direction and then using the 'bk' command, start to draw the branch at the right side. At this time, the branch at the right side should be half the size of that at the bottom. In addition, as shown in the branch at the bottom, the recursive call is also done for fractal forms at the end of the branch at the right side. Likewise, the moving and recursive call for branch at the left side are conducted in the same way as the two branches before. Finally, the LOGO turtle is moved to the basic position and then direction is moved to the basic direction to finish the procedure. If this is not conducted correctly, the turtle is not positioned at the basic direction when it is positioned at the recursive position at each corner of the branch. If the turtle is not positioned at the basic position, the fractal form is not produced. This procedure can be shown as follows if they are expressed in a flow chart.

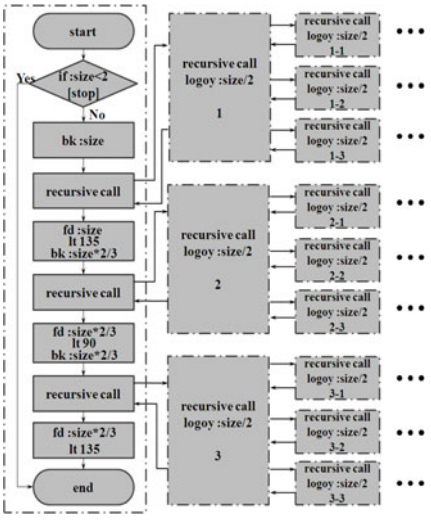


Fig. 4. Flow chart of T fractal procedure

T fractal which is drawn by using T fractal procedure is drawn from bottom through right to left as shown in Figure 5.

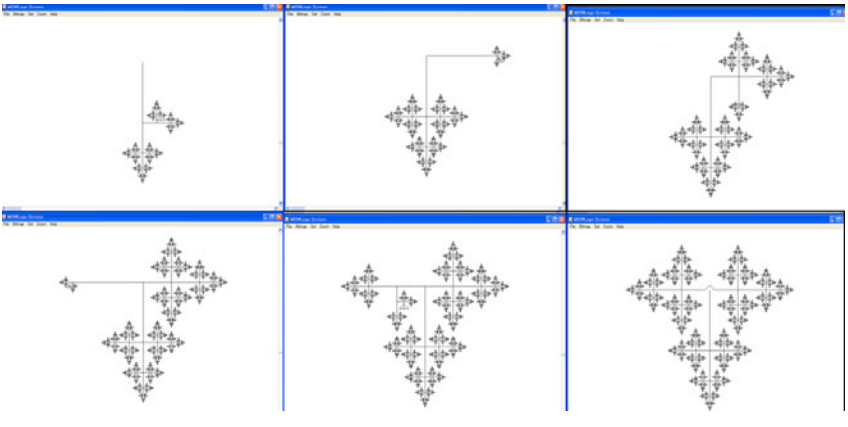


Fig. 5. Process of T fractal procedure in LOGO programming

3.3 Methods of Applying LOGO to Elementary School Computer Education

We drew up lesson plans for a 4 hour computer education course that used LOGO and fractal theory. The goal was focused not on studying the basic functions of programming but on getting the maximal effect from applying the minimum of functions. The lesson plan is shown in figure 6.

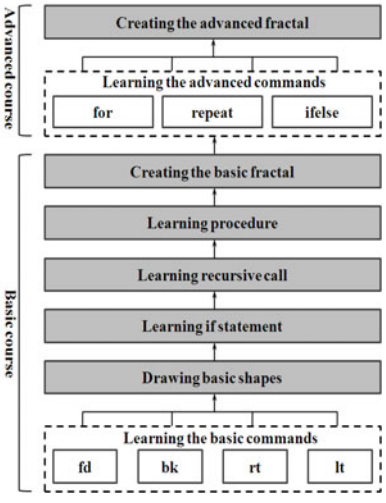


Fig. 6. Flow of Study

In the first lesson, we studied basic LOGO commands and the concept of fractals. In the second lesson, we made and studied Basic-T fractals using LOGO. In the third lesson, we made T fractals which were more difficult than Basic-T fractals. In the last lesson, we made fractal procedures using figures which were covered in the first to third lessons and that the students had learned and in their own way. The overall flow of study was composed with basic and intensified process for graded lesson.

First, the level of interest of the students has to be taken into consideration. It is possible that students will lose interest when the teacher teaches basic commands, and accordingly, the teacher has to minimize the presentation of functions. Also, the teacher can increase student interest by providing various applications which are suitable for the level of the students. Secondly, the teacher has to emphasize procedures and recursive calls because they are crucial to making fractals. Third, when the teacher teaches the T fractal, intervention has to be minimized to offer plenty of opportunities because it applies what was learned about Basic-T fractals. Lastly, advanced lessons should be given to students who get good results, and remedial education should be given to the other students.

4 Experiment and Verification of Validity

4.1 Target and Procedure of Experiment

4.1.1 Target of Experiment

The application of this study in the class was conducted with two classes of fifth graders in an elementary school. The experiment group (where the LOGO education

is applied) and the control group were made for the application of this education. To verify the homogeneity between two groups before proceeding to the application of model, the preliminary test was carried out. The result of diagnostic evaluation showed that the two groups are not significantly different from each other. In other words, two groups were found to be homogenous

Table 2. Test Unit's Person.

		A	B	C	D
Target of Experiment	Class	1	1	1	-
	People	18	18	18	-
	Time	1	1	1	0
Control Group	Class	-	-	-	1
	People	0	0	0	17
	Time	0	0	0	3

* A (General fractal education), B (LOGO programming education)
* C (Fractal + LOGO programming), C (General Math education)

4.1.2 Procedure of Experiment

The experiment group and control group were composed of the same grader students and the test was conducted at the same time. The procedure was conducted as shown in table 6 in order to understand how the suggested class model affects the students in the future.

Table 3. Test Process

Target of Experiment		C1		A1,B1	Before
	A1	→	A2	A2,B2	After
Control Group		C2		C1	Suggestion
	B1	→	B2	C2	General

4.2 Result of Experiment and Analysis of Its Effect

4.2.1 Result of Experiment

The analysis of experiment results was conducted for academic achievement, class satisfaction, and behavioral change. How much the LOGO programming education and its fusion education model based on the fractal geometrical theory affect the computer education in terms of academic achievement is shown in table 4.

Table 4. Student assessment
(Unit: %)

	Class	People	Academic achievement			
			Fractal	LOGO	General Math	Fractal + LOGO
Before	Target of Experiment	18	58	60	59	62
	Control Group	17	60	58	55	63
After	Target of Experiment	18	85	85	87	88
	Control Group	17	70	68	72	73

As shown in table 4, the students in the experiment group have shown the relatively higher academic achievement than those in control group, showing that the suggested model significantly affects the academic achievement for the students.

The satisfaction for the class is as shown in Figure 7 when three types of education are subsequently conducted three times. At first, there is not much difference in academic satisfaction as there was not much difference in input values. However, as the experiment continues, LOGO leaning model, which is conducted subsequently three times showed more satisfaction in class. This may be caused by the followings: First, the learning model using the LOGO programming could get more attention from students compared to the general mathematical education model as it reflects more interesting factors in the learning model. Second, it also helped to conduct the class in more various ways of educations and expressions and thus attracted more attention from students, causing the rise in academic satisfaction.

Third, in the process of selection among three different education models, the teacher has more chance of selecting the effective learning model and thus selectively reducing the potential conflict.

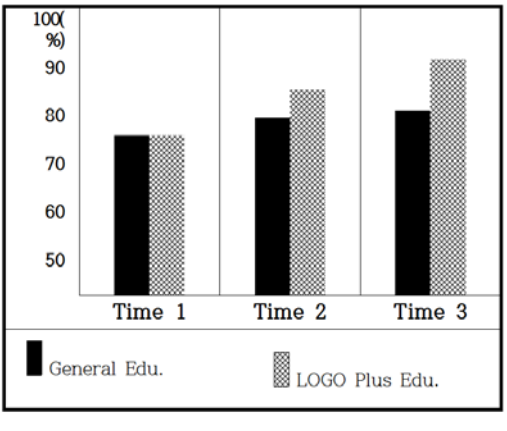


Fig. 7. Result of Class Satisfaction

5 Conclusions and Tasks

What is important in programming is not results but the process whereby they are achieved. Regardless of the result, learners will do many activities in the process of programming. During this process, problem solving and the ability to think logically will be strengthened. These advantages of learning programming also apply to elementary school students. Therefore studying programming is very important in elementary school.

We have introduced the beneficial effects of applying LOGO and fractals to elementary school computer education. One of the merits of the programming language is that the effects gained from studying it can be made stronger by connecting it to the material of other courses. So if we a variety of content that has been connected with other subjects is taught by relating it to a programming language, not only can we teach the programming language more effectively but it will also be beneficial for students as they will learn other subjects. As a result, the methods of utilizing various materials in programming language and computer education have to be studied.

References

1. Horton, B.: Integrating Logo into the Secondary Mathematics Curriculum. In: Proceedings of LOGO and Mathematics Education Conference, vol. (5) (1991)
2. Sacristan, A.I.: Exploring infinite processes through Logo programming activities of recursive and fractal figures. In: EUROLOGO Conference, vol. 10 (2005)
3. Park, N., Song, Y., Won, D.H., Kim, H.W.: Multilateral Approaches to the Mobile RFID Security Problem Using Web Service. In: Zhang, Y., Yu, G., Hwang, J., Xu, G. (eds.) APWeb 2008. LNCS, vol. 4976, pp. 331–341. Springer, Heidelberg (2008)
4. Park, N., Kwak, J., Kim, S., Won, D.H., Kim, H.W.: WIPI Mobile Platform with Secure Service for Mobile RFID Network Environment. In: Shen, H.T., Li, J., Li, M., Ni, J., Wang, W. (eds.) APWeb Workshops 2006. LNCS, vol. 3842, pp. 741–748. Springer, Heidelberg (2006)
5. Toenisson, E.: Programming language LOGO in school mathematics and teacher training. In: Proceedings of PME Conference, vol. 21(1) (1997)
6. Freiermuth, K., Hromkovič, J., Steffen, B.: Creating and Testing Textbooks for Secondary Schools - An Example: Programming in LOGO. In: Mittermeir, R.T., Sysło, M.M. (eds.) ISSEP 2008. LNCS, vol. 5090, pp. 216–228. Springer, Heidelberg (2008)
7. Park, N., Kim, H.W., Kim, S., Won, D.H.: Open Location-Based Service Using Secure Middleware Infrastructure in Web Services. In: Gervasi, O., Gavrilova, M.L., Kumar, V., Laganá, A., Lee, H.P., Mun, Y., Taniar, D., Tan, C.J.K. (eds.) ICCSA 2005, Part II. LNCS, vol. 3481, pp. 1146–1155. Springer, Heidelberg (2005)
8. Park, N., Kim, S., Won, D.: Privacy Preserving Enhanced Service Mechanism in Mobile RFID Network. ASC, vol. 43, pp. 151–156. Springer, Heidelberg (2007)
9. Park, N.: Security Scheme for Managing a Large Quantity of Individual Information in RFID Environment. In: Zhu, R., Zhang, Y., Liu, B., Liu, C. (eds.) ICICA 2010, Part II. CCIS, vol. 106, pp. 72–79. Springer, Heidelberg (2010)

10. Park, N., Song, Y.: Secure RFID Application Data Management Using All-or-Nothing Transform Encryption. In: Pandurangan, G., Anil Kumar, V.S., Ming, G., Liu, Y., Li, Y. (eds.) WASA 2010. LNCS, vol. 6221, pp. 245–252. Springer, Heidelberg (2010)
11. Park, N.: Secure UHF/HF Dual-Band RFID: Strategic Framework Approaches and Application Solutions. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 488–496. Springer, Heidelberg (2011)
12. Park, N.: Secure Data Access Control Scheme Using Type-Based Re-Encryption in Cloud Environment. In: Katarzyniak, R., Chiu, T.-F., Hong, C.-F., Nguyen, N.T. (eds.) Semantic Methods. SCI, vol. 381, pp. 319–327. Springer, Heidelberg (2011)
13. Park, N., Song, Y.: AONT Encryption Based Application Data Management in Mobile RFID Environment. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part II. LNCS (LNAI), vol. 6422, pp. 142–152. Springer, Heidelberg (2010)

Secure RFID Personal Data Management Using Privacy Reference Profile

Namje Park[1,*,**], Kwangwoo Lee[2], Sangkeun Yoo[3], Junseob Lee[3], Youngwoon Kim[3], and Hyoungjun Kim[3]

[1] Department of Computer Education, Teachers College, Jeju National University, 61 Iljudong-ro, Jeju-si, Jeju-do, 690-781, Korea
namjepark@jejunu.ac.kr
[2] Information Security Group, School of Information and Communication Engineering, Sungkyunkwan University, 300 Cheoncheon-dong, Jangan-gu, Suwon, Gyeonggi-do 440-746, Korea
kwlee@security.re.kr
[3] u-Infra Standards Reaserch Team, Standards Research Center (SRC), Electronics and Telecommunications Research Institute (ETRI), 161 Gajeong-Dong, Yuseong-Gu, Daejeon, 305-700, Korea
{lobbi,juns,qkim,khj}@etri.re.kr

Abstract. We propose a secure framework for mobile-phone based RFID service using personal privacy policy based access control for personalized ultra-high frequency (UHF) tags employing the Electronic Product Code (EPC). The framework, called mobile RPS, has dynamic capabilities that extend upon extent trust-building service mechanisms for RFID systems. This new technology aims to provide absolute confidentiality with only basic tags.

Keywords: Mobile RFID, RFID, Profile, Privacy, Security.

1 Introduction

The typical architecture of an RFID system, as defined by EPCglobal, comprises tags embedded or attached to objects, tag readers that read tag information, and a backend Information Services (IS) server that provides the required information. The tag reader can be designed to be portable or handheld, which allows for several possible applications. While RFID is most commonly used in business-to-business (B2B) commerce for managing supply channels, distribution, and logistics, there is also growing interest in the integration of tag readers with mobile phones, allowing individuals to collect and use tag data, such as in business-to-customer (B2C) marketing. And although current implementations have been limited to movie

* This research was supported by the ICT Standardization program of MKE(The Ministry of Knowledge Economy).
** Corresponding author(namjepark@jejunu.ac.kr).

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 268–276, 2011.

promotions and museums, where information security is not a major concern, continued development will see more frequent adoption in such fields as retail, medical care, and electrical drafts, where security and privacy are indispensable.

Researchers have developed many techniques to address the various security flaws in RFID systems, the simplest being to "kill" tags before they are in the hands of the user. However, because these low-cost tags have numerous applications, the user may want the tag to remain active. As one solution, Rieback *et al.* proposed the *RFID Guardian* system, which relies on a strong proxy device—a mobile phone or PDA—to mediate access by an external reader to tags for auditing of scans and tags, key management, access control, and user authentication. The *RFID Enhancer Proxy* (REP) proposed by Juels *et al.* requires the use of a similar higher-computing-power proxy device but provides better tag acquisition and ownership transfer. Kim *et al.'s Mobile Agent for RFID Privacy Protection* (MARP) is a further development that aims to provide high-level privacy by employing a public key center that manages keys for the readers, tags, server, and proxy. Kim *et al.* also recently proposed a scheme suitable for mobile-phone-based reader systems, but the method only provides reader-based authentication, which is not sufficient.

Here we propose a secure framework for mobile-phone-based RFID services using personal privacy-policy-based access control for personalized ultra-high frequency (UHF) tags employing the Electronic Product Code (EPC). The framework, called M-RPS, has dynamic capabilities that extend upon extent trust-building service mechanisms for RFID systems. This new technology aims to provide absolute confidentiality with only basic tags.

2 Overview of M-RPS Service System

Our policy-based system allows users to directly control all the information connected to tags through the security service upon purchase of any tagged items. Thus, based on the profiles, users can decide what tag information should be disclosed to whom and to what extent. The greatest strength of this mechanism is that it does not preventing all access to the information, but allows limited access to authorized persons. The user can be kept informed over the Internet of the state of compliance in connection with user-defined profiles, and analysis of system logs by the inspection function and allows identification of the fault source when a problem is reported. The main services provided for such purposes include the preparation and management of privacy policies, information access control according to set privacy policy, indication of status of operations specified by the owner, and monitoring of privacy through log management. Privacy protection in mobile RFID services requires protection systems that guarantee confidentiality and the integrity of private information on the network and ensure authorization of entities. There must be a means to provide detailed access control mechanisms that can manage object information, log data, and personal information by user group and communicate with the M-RPS systems through secure communication paths. These systems must provide auditing functions with stronger security policy-based privacy protection for each individual user defined in the RPS

system, and there must also be a mechanism to negotiate privacy policies with the mobile RFID readers to prevent them from gathering personal information.

3 Mobile RFID Privacy Service System

3.1 Major Functions of M-RPS

The proposed profile-based privacy protection service incorporates the following six functions:

1) Access control: An access control function may be used to authenticate the identity of the owner (or application service provider) and authorize access to information resources according to the privacy policies. However, this component should preferably be implemented on the service-side, since the application service provider's server should control access to all information resources and provide services according to the owner-defined profile. The service-side system must be able to deduce whether a request for access to a certain service or information should be granted on the basis of the owner-defined privacy profile, which is a formatted set of privacy protection rules and policies.

2) Registration: The application service provider (service-side system) and the user (user-side system) must register process with the M-RPS system. The M-RPS system provides a default privacy profile to the service-side system. The default privacy profile can be created by the privacy profile management functionality.

3) Privacy profile management: This is a core function of the service, involving the establishment and management of the owner's (or default) privacy profile. The system should create and manage a default profile for each service and the owner-defined profile from the registration process. Hence, this privacy profile can be sent to the service-side system when requested.

4) Privacy enhanced log management: This function performs secure event logging based on the privacy profile. Furthermore, it may be used for the detection and analysis of privacy violations in the collected event logs.

5) Obligation notification: The results of the obligation that should be performed by the service-side system can be notified to the owner via email, mobile phone text message, and so on. The service-side system should notify the RPS system of the obligation result, and the RPS system should notify the owner of the obligation result.

6) ID code refreshment: This function provides a mechanism that can refresh the ID code of a tag, and the user-side system should write this refreshed ID code to the tag. Furthermore, the service-side system should reestablish the relationship between the information and the new ID code. Hence, the ID code refreshment function seems to satisfy the ID code requirement for being untraceable.

3.2 Privacy Protection Levels

Privacy protection levels are defined by the RFID tag owner to indicate the degree of protection to be provided to different sets of data linked with a tag. In other words, if information is of a high privacy protection level, a third party may not be allowed to access this information.

1) Basic guidelines for privacy levels

Figure 1 presents an example of how the privacy levels are classified and how each level is applied. The privacy level can be set from 0, with virtually no privacy protection, to 10, where tags are killed (the level is not used here). These levels are classified into low (1–3), medium (4–6), and high (7–9.) for each application service. It is recommended that the privacy protection system should support these levels to ensure compatibility with the M-RPS system. In other words, the privacy platform has three groups of privacy protection levels from 1 to 10. The default privacy level is applied to the tag and the M-RPS system.

Object Infomation	Privacy Level: *Low Level* 1	2	3	*Medium Level* 4	5	6	*High Level* 7	8	9
Object Category	O	O	O	O	O	O	O	O	X
Object Name	O	O	O	O	O	O	O	O	X
Object Code	O	O	O	O	O	O	O	X	X
Object History	O	O	O	O	O	O	X	X	X
Price	O	O	O	O	O	X	X	X	X
Distribution Information	O	O	O	O	X	X	X	X	X
Object Description	O	O	O	X	X	X	X	X	X
Owner ID	O	O	X	X	X	X	X	X	X
Owner Account	O	X	X	X	X	X	X	X	X
Owner Personal	O	X	X	X	X	X	X	X	X

(X: Not to be disclosed, O: To be disclosed)

Fig. 1. Default privacy protection level (upon access by a third party in the logistics industry)

– Low Level (Open): Low levels refer to levels where privacy is least protected. Most mobile RFID terminals can access the system and related information, including personal information. Low levels are allowed only for those who are reliable.
– Medium Level (Object Information and History): Reliable individuals or nonprivate information are assigned medium levels, where parts of the information are not protected because some security keys are disclosed and the disclosed information does not affect security.
– High Level (Part of Object Information and Object Category): Access to high level information is not guaranteed, and all access is controlled. Only limited information such as object names or object categories are allowed in high levels. For example, at a high level, the mobile RFID service is sensitive to privacy flaws and the owner allows the least information to be exposed to third parties.

2) Privacy level data structure on RFID tag

The privacy protection level data is stored in the user data area of the tag, which is defined in the EPC Gen2 Data Format standard. The privacy protection level has TYPE, LENGTH, and VALUE in the TLV (TYPE-LENGTH-VALUE) structure, as shown below.

– TYPE, Ex.) TYPE CODE 13 = Privacy Grade

	TYPE CODE							
Bit	7	6	5	4	3	2	1	0
	0	0	0	0	1	1	0	1

Fig. 2. Privacy protection TYPE CODE field

– LENGTH : The privacy protection level stored in the user data area of the RFID is an integer, and the highest level is 10. The LENGTH field is 8 bits. The LENGTH field has been set as 110 and is of 8 bits as shown below.

	LENGTH							
Bit	7	6	5	4	3	2	1	0
	0	0	0	0	0	0	0	1
	1_{10}							

Fig. 3. Privacy protection LENGTH field

– VALUE, VALUE = INTEGER (VALUE TYPE = INTEGER), VALUE = 0: Product or service that does not require privacy protection, VALUE = 0 or higher: Product or service that requires privacy protection of level 1 or higher level.

	VALUE							
Bit	7	6	5	4	3	2	1	0
	0	0	0	0	0	0	0	0

Fig. 4. Privacy protection VALUE field

The maximum privacy level depends on the application services and can be from 1 to 9. The privacy level follows the standard for each application service, and if there is no standard, the privacy level is determined based on the results of a privacy impact assessment.

3.3 Profiling Service Mechanism

The proposed RFID privacy protection management system for mobile RFID service with privacy is shown in figure 5.

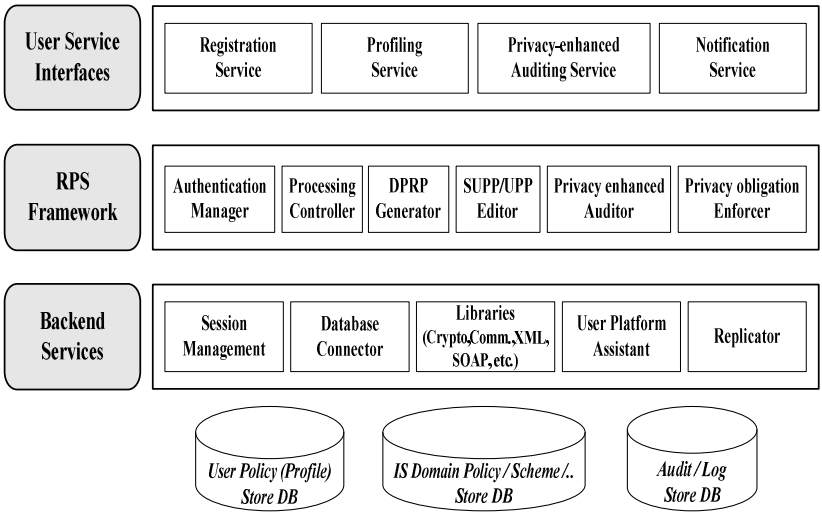


Fig. 5. System architecture for secure mobile RFID service

The main features of this service mechanism are privacy policy establishment and management, access control, obligation result notification service, and privacy audit service by audit log management. The personal privacy protection processes using the above functions are as follows.: Information is stored in a Privacy Reference List (PRL) and Privacy Reference Profile (PRP). The PRL is a list of personal information, product information, distribution information, etc., organized according to field (finance, trade, etc.). The PRP is the allotted profile for the PRL items, which is decided through effect estimation by the concerned authority (e.g., financial profile, trade profile, medical care profile, etc.). Then registration with the M-RPS and the generation of the Default Privacy Reference Profile (DPRP) occur, where each company may register and to provide their OIS schema. M-RPS generates the DPRP by matching the privacy level to each attribute of the OIS schema, which is a list of items tracked by the company from a certain service group.

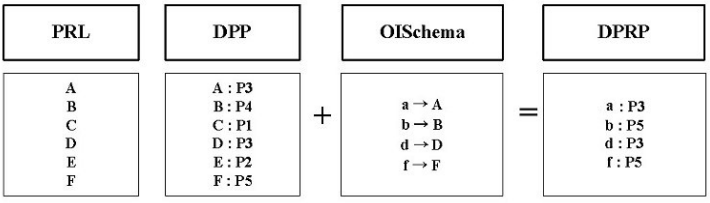


Fig. 6. Generation mechanism for DPRP

The owner's profile in the M-RPS system is generated as follows: After gaining access to the M-RPS Web server to set up an access group profile and generate the owner profile, each profile is generated in XML (eXtensible Markup Language) form by M-RPS and sent to every service provider. When the owner sets a privacy level (L1–9) for each item in the service group, an Owner-Defined Privacy Policy (OPP) for information will be generated, which in turn affords the individual Owner-Defined Privacy Reference Profile (OPRP) for service providers according to their OIS schema included in the service group. The owner must then set a single profile input for each service (to reduce the workload), as shown in Appendix A. The owner registers cellular phone numbers of those who are allowed access at every level (L1–5), and M-RPS generates a security token indicating approval, which is then sent to every service provider together with the Profile for Access Group (PAG).

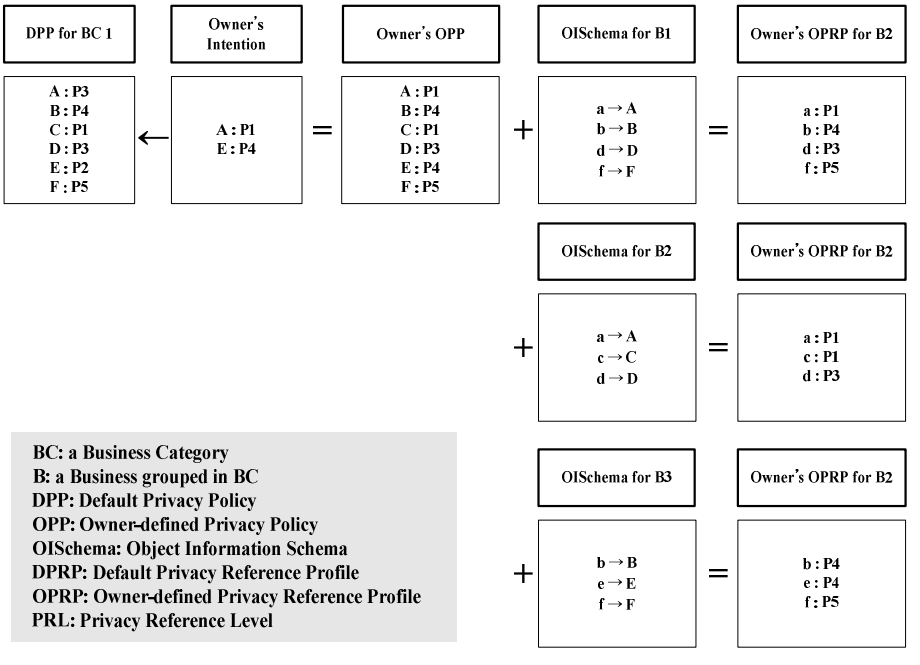


Fig. 7. Generation of OPRP

The following are performed in order by mobile RPS.

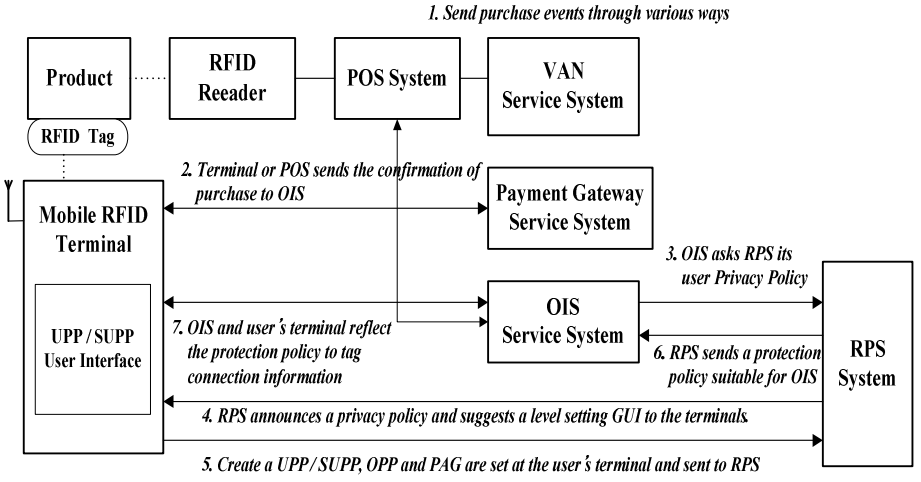


Fig. 8. Processes of M-RPS system

Purchase events are communicated through various ways. For purchase via a point-of-sale (POS) management system, owner information (cellular phone number) is sent to the OIS, and either terminal or POS sends the confirmation of purchase to the OIS, which in turn changes information in connection with the information on user. Then the OIS asks M-RPS for its user privacy policy. M-RPS announces its privacy policy and suggests a level setting GUI to the reader. OPP and PAG are set at the user's end and sent to M-RPS, which then sends a protection policy suitable for the OIS. The OIS and reader reflect the protection policy and receive only information satisfying the policy.

4 Conclusion

Mobile RFID readers are being actively researched and developed throughout the world, and more efforts are underway for the development of related service technologies. Though legal and institutional systems endeavor to protect privacy and encourage data protection, the science and engineering world must also provide suitable technologies. Seemingly, there are and will be no perfect security/privacy protection methods. The technologies proposed in this paper, however, would contribute to the development of secure and reliable RFID systems.

Acknowledgments. This paper is extended from a Ph.D. thesis paper presented at Sungkyunkwan University, South Korea (2008). The author is deeply grateful to the anonymous reviewers for their valuable suggestions and comments on the first version of this paper.

References

1. Mobile RFID Forum of Korea: Mobile RFID Privacy Protection Framework (Framework for Privacy Protection of Mobile RFID Services). MRFS-4-08. Standard Paper (2006)
2. Namje Park: Implementation of Terminal Middleware Platform for Mobile RFID computing. International Journal of Ad Hoc and Ubiquitous Computing (2011)
3. Park, N., Song, Y., Won, D.H., Kim, H.W.: Multilateral Approaches to the Mobile RFID Security Problem Using Web Service. In: Zhang, Y., Yu, G., Bertino, E., Xu, G. (eds.) APWeb 2008. LNCS, vol. 4976, pp. 331–341. Springer, Heidelberg (2008)
4. Park, N., Kwak, J., Kim, S., Won, D.H., Kim, H.W.: WIPI Mobile Platform with Secure Service for Mobile RFID Network Environment. In: Shen, H.T., Li, J., Li, M., Ni, J., Wang, W. (eds.) APWeb Workshops 2006. LNCS, vol. 3842, pp. 741–748. Springer, Heidelberg (2006)
5. Park, N., Kim, H.W., Kim, S., Won, D.H.: Open Location-Based Service Using Secure Middleware Infrastructure in Web Services. In: Gervasi, O., Gavrilova, M.L., Kumar, V., Laganá, A., Lee, H.P., Mun, Y., Taniar, D., Tan, C.J.K. (eds.) ICCSA 2005, Part II. LNCS, vol. 3481, pp. 1146–1155. Springer, Heidelberg (2005)
6. Park, N.: Security Scheme for Managing a Large Quantity of Individual Information in RFID Environment. In: Zhu, R., Zhang, Y., Liu, B., Liu, C. (eds.) ICICA 2010, Part II. CCIS, vol. 106, pp. 72–79. Springer, Heidelberg (2010)
7. Park, N.: Secure UHF/HF Dual-Band RFID: Strategic Framework Approaches and Application Solutions. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 488–496. Springer, Heidelberg (2011)
8. Park, N., Kim, S., Won, D.: Privacy Preserving Enhanced Service Mechanism in Mobile RFID Network. ASC, vol. 43, pp. 151–156. Springer, Heidelberg (2007)
9. Park, N., Song, Y.: Secure RFID Application Data Management Using All-or-Nothing Transform Encryption. In: Pandurangan, G., Anil Kumar, V.S., Ming, G., Liu, Y., Li, Y. (eds.) WASA 2010. LNCS, vol. 6221, pp. 245–252. Springer, Heidelberg (2010)
10. Tsuji, T., Kouno, S., Noguchi, J., Iguchi, M., Misu, N., Kawamura, M.: Asset management solution based on RFID. NEC Journal of Advanced Technology 1(3), 188–193 (2004)
11. Chae, J., Oh, S.: Information Report on Mobile RFID in Korea. ISO/IEC JTC 1/SC 31/WG 4 N 0922, Information paper, ISO/IEC JTC 1 SC 31 WG4 SG 5 (2005)
12. Sarma, S.E., Weis, S.A., Engels, D.W.: RFID systems, security and privacy implications. Technical Report MIT-AUTOID-WH-014, AutoID Center. MIT (2002)
13. Park, W., Lee, B.: Proposal for participating in the Correspondence Group on RFID in ITU-T. Information Paper. ASTAP Forum (2004)
14. MRF Forum: WIPI C API Standard for Mobile RFID Reader (2005)
15. MRF Forum: WIPI Network APIs for Mobile RFID Services (2005)
16. Park, N., Kim, S., Won, D.H., Kim, H.W.: Security Analysis and Implementation Leveraging Globally Networked RFIDs. In: Cuenca, P., Orozco-Barbosa, L. (eds.) PWC 2006. LNCS, vol. 4217, pp. 494–505. Springer, Heidelberg (2006)
17. Lee, J., Kim, H.: RFID Code Structure and Tag Data Structure for Mobile RFID Services in Korea. In: Proceedings of ICACT (2006)
18. Chug, B., et al.: Proposal for the study on a security framework for mobile RFID applications as a new work item on mobile security. ITU-T, COM17D116E, Q9/17, Contribution 116, Geneva (2005)
19. Kim, Y., Lee, J., Yoo, S., Kim, H.: A Network Reference Model for B2C RFID Applications. In: Proceedings of ICACT (2006)

A Study on the Secure Home Healthcare Wireless Service

Changwhan Lee[1], Dongho Won[1], and Namje Park[2,*]

[1] Information Security Group, Sungkyunkwan University,
Suwon, Korea
{chlee,dhwon}@security.re.kr
[2] Department of Computer Education Teachers College,
Jeju National University, Korea
namjepark@jejunu.ac.kr

Abstract. Home healthcare has been developed to address personal healthcare issues and to make use of advances in IT technology. Home healthcare users can check their physical information by means of many healthcare sensors and simultaneously receive remote treatment. Hence, when the functions of home healthcare are broadened, it will be possible for a patient and a medical team to access medical information or treatment using wireless communication. However, this broadening of home healthcare will reveal vulnerabilities that cannot now be predicted. Therefore, in this paper, we analysis the vulnerabilities of home healthcare wireless services, and introduce wireless security devices that will defend against these vulnerabilities.

Keywords: home healthcare service, wireless security devices.

1 Introduction

There has recently been research into home healthcare because of advances in IT technology and the increased interest in personal healthcare. Home healthcare is an IT convergence technology that combines devices and/or services with information technology. In home healthcare, the patient can monitor his physical condition and receive medical treatment from a remote location by creating a link between the patient and the hospital. The basic home healthcare component is a sensor device, such as an electrocardiogram, which monitors a patient's specific physical condition, such as the heart function or the blood sugar. This physical information is then transmitted to a monitoring device, such as a medical PC. The physical information collected from sensor devices is also transmitted to the server of the hospital that holds treatment information so that remote medical treatment can be provided. Throughout this process, home healthcare manages very important information which is directly related to the patient's privacy and the patient's life. However, wireless communication technology has developed to the point where patients and a medical team can use wireless devices to access the medical PC or treatment information database. Therefore, in this paper, we classify home healthcare devices according to

* Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 277–284, 2011.

the role they play in the service, and specify devices which support wireless communication. Furthermore, we discuss the vulnerabilities that result from using wireless communication in home healthcare, and investigate wireless security devices which can make parts of a wireless home healthcare service more secure.

In Section 2 of this paper, we briefly describe the background of home healthcare and discuss treatment information. In Section 3, we categorize the known vulnerabilities in home healthcare, and we introduce wireless security devices for home healthcare that address the vulnerability issues we discussed. Finally, we summarize and conclude our research in Section 4.

2 Overview of Home Healthcare

In this section, we provide an overview of home healthcare.

2.1 Overview of Home Healthcare

Home healthcare is a next generation medical service in which patients who use home healthcare are not subject to time and space restrictions when receiving medical treatment. Components of home healthcare have been classified into three parts. The first part is the sensor devices, which are devices that sense a patient's specific physical condition, blood sugar, blood pressure, heartbeat, and oxygen saturation, and transmit information about these physical conditions to a medical PC. The second part is the medical PC which collects and monitors information. Patients can monitor the received physical condition information on the medical PC, and the medical PC transmits the received information to a medical information server which is located in a remote hospital. The third part is the medical information server, which analyzes information transmitted from the medical PC, and a medical team may administer medical treatment to patients using the analyzed information. All data is stored in a database on the medical information server. After saving the data, a medical team or patients can access the database when data is needed about a patient. The components of home healthcare are shown in Figure 1, and each component is described in table 1 below [1].

Table 1. Components of home healthcare service

Classification	Description
Physical condition sensor	-Collect patient's physical condition information -Transmit sensing information to monitoring and control device - Based on Zigbee communication devices
Monitoring and control	- Receive information from sensor devices -Transmit collected information to medical information server in hospital -Located in home and patient check his physical condition -Receive information from monitoring and control device
Medical information server	- Analyze received medical information - Notice analyzed information to medical team - Store medical information in database

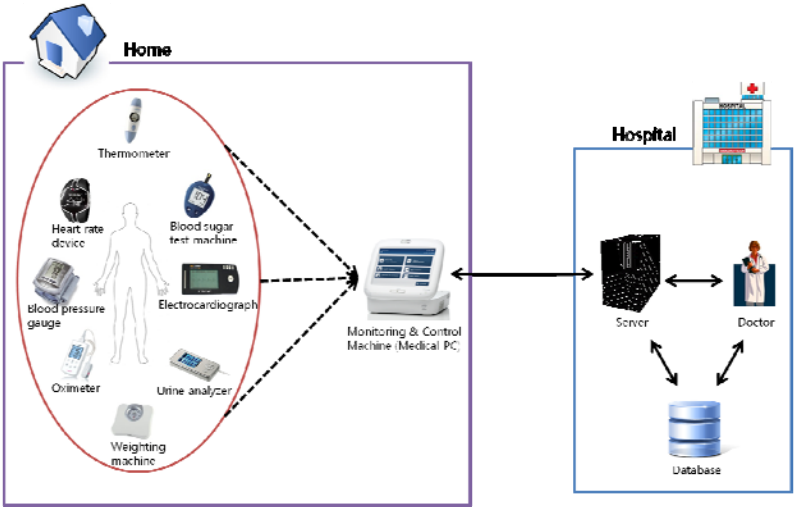


Fig. 1. Structure of home healthcare service

2.2 Treatment Information in Home Healthcare Service

As we can see in Figure 1, home healthcare consists of the above three parts. Each part handles private information and medical information that is directly connected to the life of the patient. If this information were to be leaked, the patient's privacy would be violated and a threat to the life of the patient may occur. In this section, we discuss how the devices of each part handle the information. Table 2 is a description of the treatment information handled by home healthcare components.

Table 2. Treatment information in home healthcare service [2]

Classification	Device	Description	Treatment information
Physical condition sensor	Blood sugar	Sensing and measuring patient's blood sugar level	Patient's blood sugar level
	Blood pressure	Sensing and measuring patient's blood pressure level	Patient's blood pressure level
	Heartbeat	Sensing and measuring patient's blood heartbeat	Patient's blood heartbeat
	Oxygen saturation	Sensing and measuring patient's oxygen saturation in blood	Patient's oxygen saturation in blood

Table 2. (*continued*)

Classification	Device	Description	Treatment information
	Weighing scale	Sensing and measuring patient's weight	Patient's weight
	Thermometer	Sensing and measuring patient's temperature	Patient's temperature
	Urine analyzer	Sensing and analyzing patient's urine	Patient's urine analyzed information
Monitoring and control	Medical PC	Collecting patient's physical condition from sensor devices and monitoring physical information	physical information collected from physical sensor
Medical information server	Analysis server and stored database	Analyze and store physical condition information	Privacy information Patient's medical information

3 Proposed Devices for a Secure Home Healthcare Service

3.1 Vulnerabilities of a Wireless Home Healthcare System

As Figure 1 shows, there are two types of wireless communication services used in home healthcare. The first type is wireless communication with a medical PC in a patient's home. A patient can access the medical PC using a wireless communication service, such as Wi-Fi, and check his physical condition via the medical PC. The second type is the wireless communication service of a medical information server. The medical team and patients can access and read information stored on the database server using the wireless communication service. A database server must identify and verify a user who has access to and can read the data. However, our research into the existing home healthcare systems has revealed that the wireless communication service of home healthcare has vulnerabilities which leave it open to attack. Therefore, a security system must be applied to protect the medical information used by the service. Figure 2 and Figure 3 show the vulnerabilities that exist at each part of a wireless communication service [3][4].

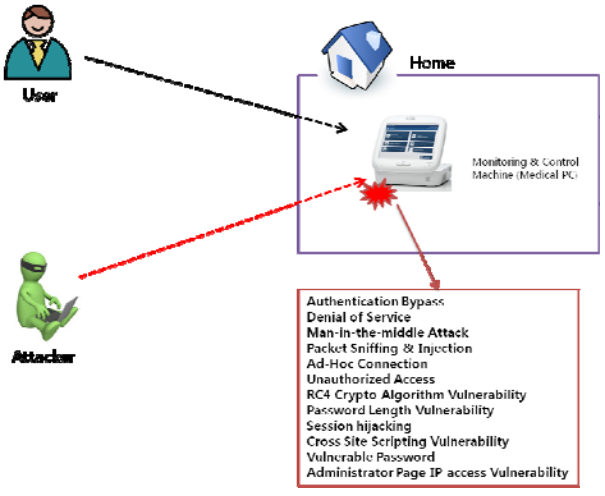


Fig. 2. Vulnerabilities and threats of home healthcare wireless service at medical PC

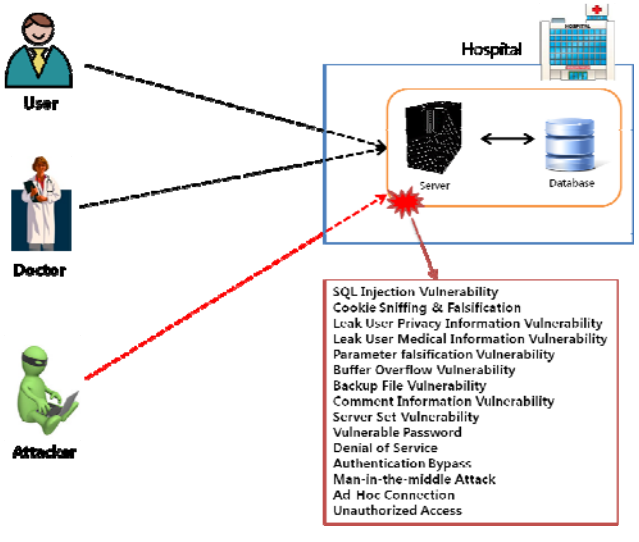


Fig. 3. Vulnerabilities and threats of home healthcare wireless service at medical server

In order to analyze the vulnerabilities of wireless communication devices, we referred to vulnerabilities analysis web sites, related journals, and portal sites that offer information about vulnerabilities. Table 3 describes vulnerabilities which may be present. Figures 2 and 3 use graphics to present some possible vulnerabilities of wireless communication devices.

Table 3. Vulnerabilities of wireless communication devices

Vulnerability	Explanation
Authentication Bypass (WEP Cracking)	Attack wireless packet security function WEP, and acquire WEP key through attack
Denial of Service	Send massive packet to AP or system for denial of service
Unauthorized Access	Unauthorized user access to system
Man-in-the-middle Attack	Malicious attacker fake valid server between users and valid server, so he acquire communication information between users and server
Packet Sniffing & Injection	Eavesdropping between valid users communication.
Ad-Hoc Connection	External unauthorized user illegal connection with internal network user using Ad-Hoc method
Miss-association AP	To leak information of company, attacker lead internal user connected other network using that wireless equipment automatic find SSID and connect property
RC4 Crypto Algorithm Vulnerability	Leak information on wireless network, using cipher algorithms vulnerability
Password Length Vulnerability	Using password length's vulnerability, attacker do social engineering hacking to get key value
Session hijacking	Attacker intercepts active session after valid user session was operated. So attacker can observe and control all operation
Cross Site Scripting Vulnerability	If user operate page, script which was transferred attacker's code would operate
SQL Injection Vulnerability	Attacker force injection SQL command to target database, and operate data leak, falsification, or administrator authentication bypass
Cookie Sniffing & Falsification	Attackers steal or control cookies in users' web browser
Leak User Privacy Information Vulnerability	User privacy information was leaked from database
Leak User Privacy Information Vulnerability	User medical information was leaked from database
Parameter falsification Vulnerability	Attacker modify normal system parameter to occur abnormal operation
Buffer Overflow Vulnerability	Attack input over buffer size of data, to purpose of operating malicious command
Backup File Vulnerability	Vulnerabilities of backup files like .bak which was generated during developed server
Comment Information Vulnerability	A comment generated during developed system possibly leak system information
Vulnerable Password	Vulnerabilities of below secure password length or possibly guess combination password
Server Set Vulnerability	Vulnerabilities of default server set value which was not changed
Administrator Page IP access Vulnerability	Attacker can unauthorized access to system because IP access control does not operate on administrator page

3.2 Wireless Security Devices That Secure a Home Healthcare Wireless Service

As we showed in section 3.1, existing home healthcare wireless services have vulnerabilities. However, some wireless security devices protect against these vulnerabilities. In this section, we discuss and apply wireless security devices which should be installed to protect the information managed by a home healthcare wireless service [5][6].

Wireless Authentication System. A wireless authentication system is composed of an authentication server and an authentication client. The authentication server allows a user to use an authentication client to get past an access point. The authentication server transmits access control policies and a master key used to secure communication between a server and a client across the access point when a server has certified a client. The authentication client has been installed in a wireless device which supports 802.1x and takes part in the certification process with the authentication server. A system manager can access the authentication server to set access control policies.

Wireless Intrusion Prevention System (WIPS). A WIPS has various functions including implementing an Intrusion Prevent System (IPS) on a wired network, network monitoring, intrusion detection, blocking, and tracing. In addition, wireless networks support such functions as detecting and capturing wireless packets on a network.

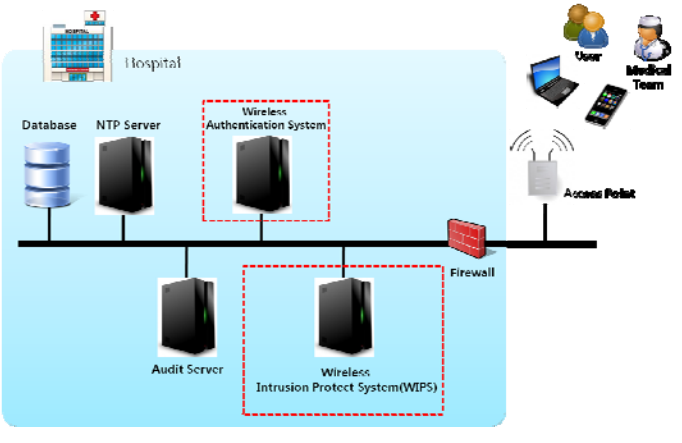


Fig. 4. Wireless security devices in home healthcare wireless service

In Figure 4, we apply the wireless authentication system and the intrusion prevention system to a home healthcare wireless communication service. These security devices protect against the vulnerabilities that were discussed in section 3.1.

4 Conclusion

Home healthcare is not subject to time and space restrictions and allows a patient to receive medical treatment in his home. As time goes by, the interest in personal healthcare will further the development of home healthcare. Simultaneously, users who use home healthcare would like the service to support a wireless communication environment. However, very important information, most of which is personal and private and is directly related to a patient's life, is processed while providing home healthcare. Therefore, this information must be protected when performing wired and wireless communication when providing home healthcare. In this paper, we analyzed the vulnerabilities which a wireless home healthcare service may have. We also researched and applied wireless security devices to protect against the vulnerabilities. When the home healthcare environment is expanded on in the future, most service components will support wireless communication. When this expansion occurs, the proper wireless security devices should be installed in order to protect a patient's information.

References

1. Lee, C., Lee, K., Kim, S., Won, D.: Analysis on Vulnerability of Home Healthcare Medical Devices and Development of Protection Profile based on Common Criteria Version 3.1. In: Proc. of CNSI 2011, International Conference on Computers, Networks, Systems, and Industrial Engineering, Jeju Island, May 23-25, pp. 240–247 (2011)
2. Jung, J.Y., Lee, J.W.: ZigBee device design and implementation for context aware U-healthcare system. In: Systems and Networks Communications, ICSNC, p. 22 (2007)
3. CVE, Internet site for vulnerability analysis, http://cve.mitre.org/
4. Andrew, A., Konstantin, V., Andrei, A.: WI-FOO The Secrets of Wireless Hacking (2004)
5. National Cyber Security Center, Wireless authentication system protection profile (2008)
6. US Government Information Assurance Directorate, Wireless Local Area Network Client Protection Profile for Basic Robustness Environment (2007)
7. Nam, J., Kim, S., Won, D.H.: Secure Group Communications Over Combined wired and Wireless Networks. In: Katsikas, S.K., López, J., Pernul, G. (eds.) TrustBus 2005. LNCS, vol. 3592, pp. 90–99. Springer, Heidelberg (2005)
8. Choi, H.-K., Shin, J.: Simulation Framework for Wireless Internet Access Networks. In: Sunderam, V.S., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2005, Part II. LNCS, vol. 3515, pp. 453–460. Springer, Heidelberg (2005)
9. Park, N., Kwak, J., Kim, S., Won, D.H., Kim, H.W.: WIPI Mobile Platform with Secure Service for Mobile RFID Network Environment. In: Shen, H.T., Li, J., Li, M., Ni, J., Wang, W. (eds.) APWeb Workshops 2006. LNCS, vol. 3842, pp. 741–748. Springer, Heidelberg (2006)

Cryptanalysis of the User Authentication Scheme with Anonymity*

Woongryul Jeon[1], Jeeyeon Kim[1],
Junghyun Nam[2], Youngsook Lee[3], and Dongho Won[1,**]

[1] School of Information and Communication Engineering, Sungkyunkwan University, 300 Cheoncheon-dong, Jangan-gu, Suwon-si, Gyeonggi-do, 440-746, Korea
[2] Department of Computer Engineering, Konkuk University, 322 Danwol-dong, Chungju-si, Chungcheongbuk-do, 380-701, Korea
[3] Department of Cyber Investigation Police, Howon University, 727 Weolha-li, Impi-Myeon, Gunsan-si, Jeonrabuk-do, 573-718, Korea
{wrjeon,dhwon}@security.re.kr, jeeyeonkim@paran.com, jhnam@kku.ac.kr, ysooklee@howon.ac.kr

Abstract. Nowadays, wireless communications using mobile device are growing rapidly. The most advantage of wireless communication is that user can transfer various information to anywhere at anytime using mobile device. However, it is clear that security of the wireless communications is more complex than wired owing to the openness of the wireless network. Thus, authentication has become the most important issue in the wireless communication. Furthermore, owing to the fact that location information shows user's life style, user anonymity becomes also significant issue. Recently, in 2011, Kang et al. pointed out that Wu et al.'s scheme has some security vulnerabilities and proposed an improved scheme. However, in this paper, we discuss that Kang et al.'s scheme is still vulnerable to active adversary.

Keywords: wireless communication, anonymity, user authentication, security.

1 Introduction

Nowadays, wireless communications using mobile device are growing rapidly. The most advantage of wireless communication is that user can transfer various information to anywhere at anytime using mobile device. However, it is clear that security of the wireless communications is more complex than wired owing to the openness of the wireless network. [1].

Recently, user anonymity also becomes significant issue in the mobile device. Most of the mobile devices have GPS sensor, and many services are based on

* This research was supported by Basic Science Research Program thorough the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2011-0004751), This work was supported by Priority Research Centers Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2011-0018397).
** Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 285–292, 2011.

location information. Thus, the mobile device, which records location information, reveals life style of the user. Therefore, user anonymity becomes significant issue to protect privacy from the location information.

In 2004, Zhu and Ma proposed an authentication scheme with user anonymity for wireless communications [2]. However, they do not provide user anonymity, actually. After Zhu and Ma, many subsequent studies have found and correct the security weaknesses [2]-[13]. In 2006, Lee et al. pointed out that Zhu et al.'s scheme is susceptible to forgery attack and proposed an improved scheme [3]. In 2008, Wu et al. revealed that both authentication scheme do not provide anonymity as claimed and proposed a modified scheme [4]. Recently, in 2011, Kang et al. pointed out that Wu et al.'s scheme has some security problems and proposed an improved scheme [5]. However, in this paper, we show that Kang et al.'s scheme is still susceptible to active adversary.

The remainder of this paper is organized as follows. In section 2, we briefly review Kang et al.'s scheme and discuss security analysis of the Kang et al.'s scheme, in section 3. Conclusions are given in section 4.

2 Review of Kang et al.'s Scheme

In this section, we review Kang et al.'s scheme. Following table 1 shows all notations used throughout in this paper.

- HA : Home agent of a mobile user.
- FA : Foreign agent of the network.
- MU : Mobile user.
- ID_A : Identity of an entity A.
- PW_A : Password of an entity A.
- T_A : Timestamp generated by an entity A.
- $Cert_A$: Certificate of an entity A.
- N : Strong secret key of HA.
- $(M)_K$: Encryption of a message M using a symmetric key K.
- $E_K(M)$: Encryption of a message M using an asymmetric key K.
- S_A : Secret key of an entity A.
- P_A : Public key of an entity A.
- $h(\cdot)$: Cryptographic one-way function.
- $\oplus$: Bitwise exclusive-or(XOR) operation.

As previous researches, Kang et al.'s scheme consists of three phases: initial phase, first phase and second phase. Now, we review Kang et al.'s scheme by phases.

2.1 Initial Phase

To start wireless communications, mobile user has to register itself to HA. The initial phase is invoked when mobile user MU register itself to HA. Following steps and fig.1 show the initial phase. The statement $A \rightarrow B$:M denotes that A sends a message M to B.

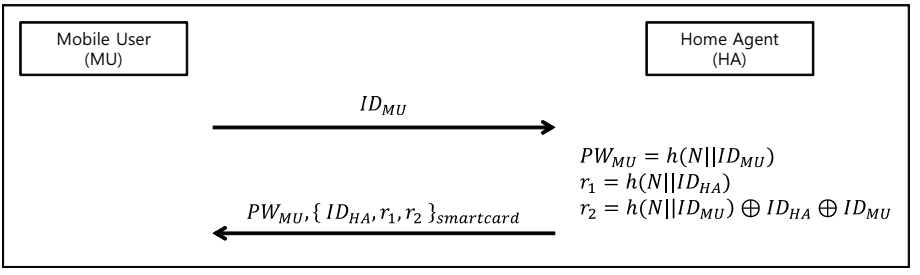


Fig. 1. Initial phase of Kang et al.'s scheme

1. $MU \rightarrow HA : ID_{MU}$
   MU choose his/her ID freely, and then sends it to HA via a secure channel.
2. $HA \rightarrow MU : PW_{MU}$, $\{ID_{HA}, r_1, r_2, h(\cdot)\}_{Smartcard}$
   Upon receiving the ID_{MU} from MU, HA computes PW_{MU}, r_1, and r_2 as follows.

   - $PW_{MU} = h(N||ID_{MU})$
   - $r_1 = h(N||ID_{HA})$
   - $r_2 = h(N||ID_{MU}) \oplus ID_{HA} \oplus ID_{MU}$

   After this computation, HA stores ID_{HA}, r_1, and r_2 into smartcard and sends it to MU with PW_{MU} via a secure channel.

2.2 First Phase

The first phase is started when MU visits foreign agent HA. Following steps and fig.2 describe the first phase.

1. $MU \rightarrow FA : n$, $(h(ID_{MU})||x_0||x)_L$, ID_{HA}, T_{MU}
   When MU enters PW_{MU} to mobile device, mobile device generates random values x_0 and x. Then mobile device computes both $n = h(T_{MU}||r_1) \oplus r_2 \oplus PW_{MU}$ and $L = h(T_{MU} \oplus PW_{MU})$. Finally, MU generates its timestamp T_{MU} and sends n, $(h(ID_{MU})||x_0||x)_L$, ID_{HA}, and T_{MU} to FA.
2. $FA \rightarrow HA : b$, n, $(h(ID_{MU})||x_0||x)_L$, T_{MU}, $Sign_{S_{FA}}$, $Cert_{FA}$, T_{FA}
   FA checks the validity of T_{MU}. If it is valid, then FA generates a random number b, and computes a digital signature $Sign_{S_{FA}} = (h(h(ID_{MU})||x_0||x)_L$, T_{MU}, $Cert_{FA}))$. Finally, FA sends b, n, $(h(ID_{MU})||x_0||x)_L$, T_{MU}, $Sign_{S_{FA}}$, $Cert_{FA}$, and T_{FA} to HA.
3. $HA \rightarrow FA : c$, W, b, $Sign_{S_{HA}}$, $Cert_{HA}$, T_{HA}
   Upon receiving the message from FA, HA verifies T_{FA} and $Cert_{FA}$. If they are valid, then HA retrieves MU's real identity ID'_{MU} as follows.

- $ID'_{MU} = n \oplus h(T_{MU}||r_1) \oplus ID_{HA} = n \oplus h(T_{MU}||h(N||ID_{HA})) \oplus ID_{HA}$
Then HA computes $L = h(T_{MU} \oplus PW_{MU})$ and decrypts $(h(ID_{MU})||x_0||x)_L$ to obtain $h(ID_{MU})$. If $h(ID'_{MU})$ is equal to $h(ID_{MU})$, then MU is a legal user of HA.
After authentication, HA generates a random number c, and then computes both $W = E_{P_{FA}}(h(h(N||ID_{MU}))||x_0||x)$ and its corresponding digital signature $Sign_{S_{HA}} = (h(b, c, W, Cert_{HA}))$. Finally, HA sends c, W, b, $Sign_{S_{HA}}$, $Cert_{HA}$, and T_{HA} to FA.

4. $FA \rightarrow MU : (TCert_{MU}||h(x_0||x))_{sk}$
Upon receiving the message from HA, FA checks T_{HA} and $Cert_{HA}$. If they are valid, then FA decrypts W and obtains $(h(h(N||ID_{MU}))||x_0||x)$. Now, FA computes a common session key $sk = h(h(h(N||ID_{MU}))||x||x_0) = h(h(PW_{MU})||x||x_0)$. Finally, FA issues a temporary certificate $TCert_{MU}$ and encrypts it with sk. Then FA sends $(TCert_{MU}||h(x_0||x))_{sk}$ to MU.
5. Upon receiving the message from FA, MU computes sk and decrypts the message. Then MU computes $h(x_0||x)$ and compare with decrypted value. If they are equal, then MU authenticates FA successfully. As a result, MU establishes sk with FA.

2.3 Second Phase

The second phase is invoked when MU visits FA at the i-th session. To enhance the efficiency of the scheme, MU sends following message to FA.

$$MU \rightarrow FA : TCert_{MU}, (x_i||TCert_{MU}||Other\ Information)_{sk_i}$$

When FA receives the message from MU, then FA verifies $TCert_{MU}$ to determine whether MU is a legal user or not. If MU is a legal user, FA decrypts $(x_i||TCert_{MU}||Other\ Information)_{sk_i}$ with current session key sk_i and obtains x_i. Then FA computes next session key $sk_{i+1} = h(h(PW_{MU})||x||x_i)$.

3 Cryptanalysis of Kang et al.'s Scheme

This section described security analysis of Kang et al.'s scheme. In short, Kang et al.'s scheme is susceptible to active adversary. Following shows general assumptions for authentication using smartcard [10],[13]-[14].

1. An adversary can have total control over the communication channel among MU, FA, and HA in the scheme. That is, an adversary can intercept, insert, delete, or modify any message in the channel.
2. An adversary can obtain or steal legal user MU's smartcard and then extract the information from it.

Now, we discuss the security vulnerabilities of Kang et al.'s scheme based on these assumptions.

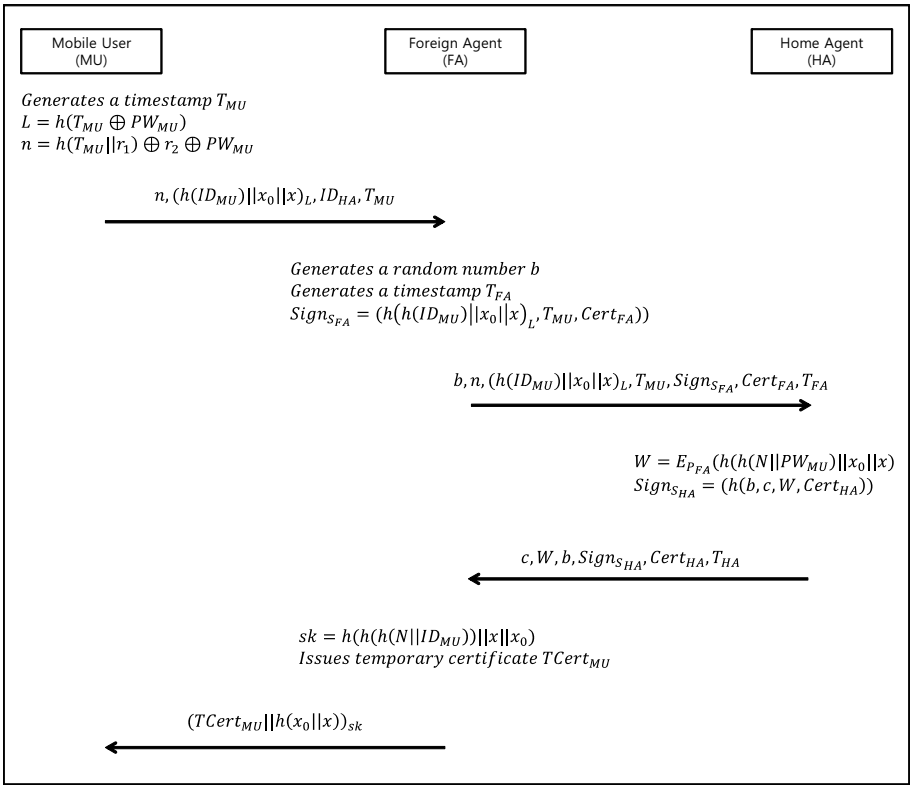


Fig. 2. First phase of Kang et al.'s scheme

3.1 Password Exposure

Based on the assumption 2, an adversary can steal MU's smartcard which contains ID_{HA}, r_1, and r_2, and an adversary also can capture both n and T_{MU} from wireless communication in the first phase, based on the assumption 1.

Now, an adversary can retrieve PW_{MU} easily as follow.

$$PW_{MU} = n \oplus h(T_{MU}||r_1) \oplus r_2$$

An adversary can perform many attacks with this retrieved PW_{MU}. With this PW_{MU}, an adversary can computes L and decrypts $(h(ID_{MU})||x_0||x)_L$ to obtain both x_0 and x. Then an adversary can computes the session key $sk = h(h(PW_{MU}||x||x_0)$. Therefore, an adversary can decrypt all messages between MU and FA.

Furthermore, an adversary also can decrypt the messages in the second phase, for example, $(x_i||TCert_{MU}||Other\ Information)_{sk_i}$. Thus, an adversary can obtain x_i. When x_i is exposed to an adversary, then he/she can compute next session key as $sk_{i+1} = h(h(PW_{MU})||x||x_i)$.

Therefore, Kang et al.'s scheme does not provide perfect forward secrecy.

3.2 User Anonymity

Likewise above section, with r_1, r_2, n and T_{MU}, an adversary also can retrieve the MU's password PW_{MU} easily as follow.

Now, an adversary can obtain MU's real identity ID_{MU} as follow.

$$ID_{MU} = n \oplus ID_{HA} \oplus h(T_{MU}||r_1)$$

When, an adversary obtain MU's password, then adversary can retrieve MU's real identity without assumption 1 as follow.

$$ID_{MU} = r_2 \oplus PW_{MU} \oplus ID_{HA}$$

Therefore, Kang et al.'s scheme does not provide user anonymity.

3.3 User Friendliness

In the Kang et al.'s scheme, MU's password is not selected by MU, but computed by HA as $h(N||ID_{MU})$. Furthermore, this password is too long for MU to remember, thus MU has to record this password for authentication. However, this can cause another security problem, than user may lost his/her record easily.

Thus, recently, most of the researches use user-friendly password. Strictly speaking, Kang et al.'s scheme is impractical.

Following table 1 shows general requirements of authentication scheme with anonymity for wireless communications [12]-[13].

Table 1. Requirements for user authentication scheme

Requirements	Kang et al.'s scheme
User friendliness	X
Without password table	O
Freely change password	X
User anonymity	X
Perfect forward secrecy	X
Password exposure	X

Therefore, we can conclude that Kang et al.'s scheme is not appropriate as authentication scheme with anonymity for wireless communication.

4 Conclusion and Future Work

Recently, user anonymity also becomes significant issue in the mobile device. Most of the mobile devices have GPS sensor, and many services are based on location information. Thus, the mobile device, which records location information, reveals life style of the user. Therefore, user anonymity becomes significant issue to protect privacy from the location information.

After Zhu and Ma, many subsequent studies were announced, but most of them failed to provide user anonymity. Recently, in 2011, Kang et al. pointed out that Wu et al.'s scheme has some security problems and proposed an improved scheme. However, in this paper, we showed that Kang et al.'s scheme is totally broken based on general assumptions.

The cause of these vulnerabilities it that secret parameter r_1 and r_2 is stored in smartcard. Although Kang et al. divided r into r_1 and r_2 to enhance security, however, on the contrary to Kang et al.'s aim, it makes worse. To deter these vulnerabilities, the entire scheme has to be re-designed and Lee and Kwon's scheme[15] would be good reference.

References

1. Nam, J., Paik, J., Kang, H., Kim, U., Won, D.: An off-line dictionary attack on a simple three-party key exchange protocol. IEEE Communication Letters, 205–207 (2009)
2. Zhu, J., Ma, J.: A new authentication scheme with anonymity for wireless environments. IEEE Trans. Consum. Electiron 50(1), 13(3), 230–234 (2009)
3. Lee, C.C., Hwang, M.S., Liao, I.E.: Security enhancement on a new authentication scheme with anonymity for wireless environments. IEEE Trans. Ind. Electron. 53(5), 1683–1687 (2006)
4. Wu, C.C., Lee, W.B., Tsaur, W.J.: A secure authentication scheme with anonymity for wireless communications. IEEE Commun. Lett. 12(10), 722–723 (2008)
5. Kang, M., Rhee, H., Choi, J.: Improved User Authentication Scheme with User Anonymity for Wireless Communications. IEICE Trans. Fundamentals E94-A(2), 860–864 (2011)
6. Zeng, P., Cao, Z., Choo, K.R., Wang, S.: On the anonymity of some authentication schemes for wireless communications. IEEE Commun. Lett. 13(3), 170–171 (2009)
7. Lee, J., Chang, J.H., Lee, D.H.: Security flaws of authentication scheme with anonymity for wireless communications. IEEE Commun. Lett. 13(5), 292–293 (2009)
8. Wei, Y., Qui, H., Hu, Y.: Security analysis of authentication scheme with anonymity for wireless environments. In: ICCT(International Conference on Communication Technology) (2006)
9. Fan, C., Chan, Y., Zhang, Z.: Robust remote authentication scheme with smart cards. Comput. Secur. 24(8), 619–628 (2005)
10. Yoon, E., Yoo, K., Ha, K.: A user friendly authentication scheme with anonymity for wireless communications. Computres and Electrical Engineering 37, 356–364 (2011)
11. Xu, J., Zhu, W.T., Feng, D.G.: An improved smart card based password authentication scheme with provable security. Computer Standards & Interfaces 31(4), 723–728 (2009)
12. Li, C.T., Lee, C.C.: A novel user authentication and privacy preserving scheme with smart cards for wireless communications, Mathematical and Computer Modelling (In press)

13. Yoon, E., Yoo, K., Ha, K.: A user friendly authentication scheme with anonymity for wireless communications. Computers and Electrical Engineering 37(3), 356–364 (2011)
14. Lee, Y., Nam, J., Kim, S., Won, D.: Two Efficient and Secure Authentication Schemes Using Smart Cards. In: Gavrilova, M.L., Gervasi, O., Kumar, V., Tan, C.J.K., Taniar, D., Laganá, A., Mun, Y., Choo, H. (eds.) ICCSA 2006, Part V. LNCS, vol. 3984, pp. 858–866. Springer, Heidelberg (2006)
15. Lee, J., Kwon, T.: Secure Authentication Scheme with Improved Anonymity for Wireless Environments. IEICE Trans. Commun. E97-B(2) (2011)

An Improved Anonymous Electronic Prescription Scheme*

Chanjoo Chung[1,3], Kwangwoo Lee[3], Jungmee Yun[2,3], and Dongho Won[3,**]

[1] Financial Supervisory Service,
#38 Yeoui-daero, Youngdeungpo-gu, Seoul, 150-743
[2] Korea Electronics Technology Institute,
#68 Yatap-dong, Bundang-gu Seongnam, Gyeonggi-do 463-816, Korea
[3] Information Security Group, School of Information and Communication Engineering,
Sungkyunkwan University, 300 Cheoncheon-dong, Jangan-gu,
Suwon, Gyeonggi-do 440-746, Korea
cjchung@fss.or.kr, {kwlee,dhwon}@security.re.kr,
yunjm@keti.re.kr

Abstract. The existing electronic prescription scheme fails to protect identities to mandates party or to health insurance. To solve this problem, we propose RSA cryptosystem based anonymous electronic prescription which is issued from university and local hospitals by authorized medical professionals. The proposed scheme protects the identity exposure of doctors and privacy of patients. The proposed approach will help national health insurance corporation to increase the transparency of national prescription system.

Keywords: Electronic prescription, Anonymity, Privacy, RSA, PKI.

1 Introduction

In 2004, Y. Yang et al.[2] proposed a description of protocols that were used for B. Lee's conventional electronic prescription scheme[1] based on proxy signature prior to their new, signer-counterfeit-proof scheme. Since the signer-counterfeit-proof electronic prescription scheme applies a digital signature which is based on discrete logarithms, the current prime factorization-based PKI should be replaced with a discrete logarithm-based PKI, if you want to apply this scheme to the real world. Look at domestic and foreign certificates used in Internet Explorer, and you can see that most of their PKIs are based on prime factorization. Another problem is that at the delegation phase for the prevention of signature forging, a larger amount of calculation is required compared with B. Lee's scheme.

To solve this problem, we propose an RSA based anonymous electronic prescription system, in which doctors at university hospitals and clinics can issue

* 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(2011-0026023).

** Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 293–300, 2011.

prescriptions electronically. This system utilizes the conventional PKI environment to ensure the anonymity of the doctor who has issued a prescription and the privacy of the patient who has been received that prescription. In addition, in this system, since the NHIC can manage electronic prescriptions on-line, dual payment of medical bills as well as dual use of the same prescription can be prevented in advance. If an electronic ID card system is established and the NHIC runs the OCS, the proposed system will contribute to strengthen the transparency of the existing health insurance service system.

2 Security Requirements for an Electronic Prescription Scheme

Table 1 shows general security requirements[3] that an electronic prescription scheme have to have and the requirements[4] that are needed for the proxy signature-based electronic prescription proposed by B. Lee.

Table 1. General security requirements for an electronic prescription

Security requirements	Description
Unforgeability	Except the creator of the prescription, no one can be able to create a legitimate prescription.
Verifiability	Anyone who uses information available by the public should be able to verify issued prescriptions.
Identifiability	It should be possible to find out the identity of the signer using the electronic prescription that has been made by the creator.
Non-repudiation	The electronic prescription creator should not be able to deny prescriptions that he or she has issued.
Prevention of abuse	A key pairs that is created for an electronic prescription should not be used for other purposes.
Confidentiality	The information on an electronic description which is being delivered should not be accessed by anyone else except those who have the right to do.
Right to choose a dispenser	Only the dispenser who has been chosen by the patient or his or her guardian should be able to read the prescription.

The anonymous electronic prescription scheme that we propose in this paper consists of two subsystems: one is used to ensure the anonymity of the doctor who issues a prescription, and the other is used to monitor which *GP*s frequently prescribe specific medicines so that the creation of an illegal tie between *GP*s and pharmacies or between *GP*s and drug manufacturers can be prevented in advance.

This system ensures both the anonymity of the *GP* who has issued a prescription and the legitimacy of the prescription. As of today, under the OCS, since paper prescriptions that are delivered to patients have the names of the *GP*s who have issued those prescriptions, which kinds of drugs are frequently chosen by which *GP*s can be easily known. For this reason, drug manufacturers are easily able to know which *GP*s frequently choose their drugs, only by collecting prescriptions that are delivered to pharmacies, and this is being seen as a social problem.

Traceability should be ensured for all electronic prescriptions so that who made a mistake can be found out when a medical dispute occurs. In addition, an electronic prescription that was already used for dispensing medicines should not be able to be reused so that abuse of medicines is prevented. Additional security requirements for the signer-counterfeit-proof electronic prescription scheme, which was proposed by Y. Yang and his researchers, are described in Table 2.

Table 2. Additional security requirements for an electronic prescription

Additional security requirements	Description
Prevention of creation of a cartel	A potential cartel of *GP*s who create prescriptions, dispensers who distribute medicines and drug manufacturers who sell medicines described on prescriptions should be prevented in advance.
Traceability	Prescriptions should provide information on who the patient was and who the *GP* was for the time when a medical dispute happens, so as to judge who made a mistake.
Disconnectivity for doctors	Electronic prescriptions that have been issued by the same doctor should ensure pseudonymity and disconnectivity, so that pharmacists cannot know who the prescriber was.
Prevention of reuse	The same prescription should not be able to be reused so that abuse of medicines is prevented.
Pseudonymity	Electronic prescriptions should ensure pseudonymity so that the *GP* and the patient conceal themselves, and such pseudonymity should be able to be canceled by a reliable organization.
Minimum data exposure	Except the case in which information on a prescription has to be uncovered, all information on electronic prescriptions should not be uncovered.

Table 3 Provides the notations that are used in this paper.

Table 3. Notations

Notation	Description
$Sig_X(Y)$	RSA-based digital signature scheme using a private key X for a message Y.
$\sigma_X(Y)$	Discrete logarithm-based digital signature scheme using a private key X for a message Y.
$veri(\cdot)$	Algorithm for the verification of digital signature.
ω_o	Delegation warranty for an electronic prescription.
$E_X(Y)$	Encryption of a message Y using a key X.
$D_X(Y)$	Decryption of a message Y using a key X.
$H(X)$	A message X as a secure, one-way hash function.
H	Result value that is created by a secure, one-way hash function.
$\leftarrow_R$	Selection of a random number in a value range R.

Table 3. (*continued*)

$A \| B$	Concatenation between message A and B.
$\Phi(X)$	Euler function of the input X.
x_A / y_A	Discrete logarithm private key and public key of A.
pk_A / sk_A	Public key and private key of A.
p_A, q_A	Large primes of A, p and q.
$RegNo$	Registration number that was given by a registration agent or a certificate institution for applying for a certificate.
e_{HA}	Healthcare authority's RSA public key exponent ($e_{HA} \cdot d_{HA} = 1 \bmod \Phi(N_{HA})$.
N_{HA}	Modular value of a healthcare authority's RSA public key and private key. ($N_{HA}=p_{HA} \cdot q_{HA}$).
d_{HA}	RSA private key exponent of a healthcare authority. ($d_{HA}=d_C \cdot d_L \bmod \varphi(N_{HA})$).
d_C / d_L	RSA private key exponent of the central healthcare authority/ local healthcare authority.
g	Primitive element on Z^*_p.
M / WN/ SN	The content/ Writer name/ Serial number of an electronic prescription.
$Cert_A$	A's certificate.
AP_A	A's anonymous electronic prescription.
TM_i	An i message that is transmitted or a Temp message.
k	1,024 bit or 2,048 bit security parameter.

3 RSA-Based Anonymous Electronic Prescription

In this section, we propose an RSA-based anonymous electronic prescription scheme which is frequently shown in domestic and foreign PKIs.

3.1 System Setting

Suppose that the patient P receives a smart card containing a certificate $Cert_P$ that includes his or her public key pair (e_P, N_P) from a healthcare authority and access to the content of the smart card can be perfectly controlled through authentication via PIN. $e_P \cdot d_P = 1 \bmod \varphi(N_P)$ is valid.

GP makes a visit to a certification authority CA or an agent of the certification authority RA and receives a registration number for the issuance of a certificate $RegNo$ [9]. Using this number, GP gets a public key pair (pk_{GP}, sk_{GP}) and sends a digital signature $Sig_{skGP}(RegNo \| pk_{GP})$, $RegNo$ and pk_{GP} to CA via a security channel to apply for a certificate. At this time, $pk_{GP}=(e_{GP}, N_{GP})$, $sk_{GP}=(d_{GP}, N_{GP})$ and $e_{GP} \cdot d_{GP}=1 \bmod \varphi(N_{GP})$ are valid. CA uses pk_{GP} to verify a digital signature $\mathrm{Sig}_{skGP}(RegNo \| pk_{GP})$ that it has been received, uses $RegNo$ to confirm the applicant, and issues a certificate $Cert_{GP}$ that will be given to GP. GP uses $Cert_{GP}$ to register himself or herself to a HA, and then he or she can log onto the electronic prescription system using the private

key of his or her $Cert_{GP}$. *HA* consists of the *CHA* and *LHA*, and the process of the issuance of a certificate $Cert_{HA}$ is the same as that of *GP*. A healthcare authority *HA*'s private key is divided into d_C and d_L that satisfy $d_{HA}=d_C \cdot d_L \bmod \varphi(N_{HA})$. d_C and d_L indicate the private keys of the *CHA* and a *LHA* for partial signature for an anonymous electronic prescription, respectively. *LHA* has a public key pair (pk_{LHA}, sk_{LHA}) for the encryption of transmitted information and provides pk_{LHA} to *CHA*.

The patient *P* brings a smart card with him or her to visit a hospital and hand over the card to a general practitioner *GP*. After examining *P*, *GP* logs onto *HA* using his or her $Cert_{GP}$ to issue an anonymous electronic prescription. At this time, a Secure Socket Layer (SSL) or Transport Layer Security (TLS) security channel is established between *GP* and *HA*.

3.2 Capacity Delegation for an Anonymous Electronic Prescription and Partial Signature Phase

GP keeps an electronic prescription writer *WN* anonymous, makes an electronic prescription *M* consist of (*WN*, *SN*, Prescription information), and input *M* in a secure, one-way hash function to calculate $h=H(M)$. *GP* selects a blind factor $b \leftarrow_R \{0,1\}^k$ to ensure anonymity for an electronic prescription, calculates $TM_1=h \cdot b^{e_{HA}} \bmod N_{HA}$ and sends TM_1 to *CHA*.

CHA uses the TM_1 that has been sent from *GP* to calculate $TM_2= TM_1^{d_C} \bmod N_{HA}$ using its private key, and store the result values in a secure database. In order to prevent the counterfeit or forging of a partial signature by *GP*, *CHA* uses *LHA*'s public key to encrypt to be like $TM_3=E_{pkLHA}(TM_2)$ before sending it to *GP*. *GP* sends $TM_4=(h, b, TM_3)$, which includes TM_3, h and a blind factor b, to *LHA*. Among the content of TM_4 that was sent by *GP*, *LHA* decrypts TM_3 using its decryption key sk_{LHA} to obtain $TM_2=D_{skLHA}(TM_3)$. For TM_2, *LHA* signs using its partial signature private key d_L and calculates $TM_5=TM_2^{d_L} \bmod N_{HA}$. *LHA* multiplies TM_6 and the inverse number of a blind factor $h=(TM_5 \cdot b^{-1})^{e_{HA}} \bmod N_{HA}$ together, and uses *HA*'s public key e_{HA} to verify if the value is identical to h of TM_4 by checking if $h=(TM_5 \cdot b^{-1})^{e_{HA}} \bmod N_{HA}$. If the two values are identical, *LHA* stores (TM_5, h) in a secure database, calculates $h^{d_{HA}} =TM_5 \cdot b^{-1} \bmod N_{HA}$ and sends the result value to *HA*. *HA* stores $h^{d_{HA}}$ that it has been received from *LHA* in database in the form of ($h^{d_{HA}}$, *valid*), and this is a legitimate electronic prescription. *LHA* sends TM_5 to *GP*.

For TM_5 that has been received from *LHA*, *GP* calculates $TM_5 \cdot b^{-1} \bmod N_{HA}$ to solve $h^{d_{HA}}$. Then, *GP* uses a public key e_{HA} of *HA* to verify if the solved $h^{d_{HA}}$ is identical to h that he or she has generated, and if the two values are identical, GP gets a proxy signature for *P*'s anonymous electronic prescription $AP_P= h^{d_{HA}} \bmod N_{HA}$. *GP* puts an electronic prescription *M* and a digital signature for the prescription AP_P in the smart card that the patient gave him or her for medical examination.

3.3 Anonymous Electronic Prescription Verification Phase

The patient *P* got back his smart card from *GP*, and goes to a drugstore and asks a pharmacist to dispense medicines in accordance with the electronic prescription.

PH gets *M* and AP_P from the patient's smart card and checks the content of the anonymous electronic prescription, and verifies the legitimacy of the digital signature in the prescription AP_P by checking if the electronic prescription *M*'s hash function $h=H(M)$ is identical to $AP_P^{e_{HA}} \bmod N_{HA}=h$ using *HA*'s public key e_{HA}.

HA uses AP_P that it has received from *PH* to search for the state information of the electronic prescription from the database and delivers it to *PH*. If the state information is valid, *PH* dispenses medicines to *P* in accordance with the anonymous electronic prescription, and sends the message "*used*" to *HA* to say that AP_P were dispensed and sold.

4 Security Analysis

Unforgeability: Since an electronic prescription that is issued by the general practitioner *GP* after examining his or her patient comes with an RSA digital signature created by *HA*'s private key d_{HA}, all entities associated with the issuance of the prescription-- *GP*, *HA* and *PH* are able to be verified. Therefore, it is impossible to forge an electronic prescription because only a legitimate *HA* can create an electronic prescription. *GP* who has the right to write a prescription can log onto the system only by putting a digital signature using his or her certificate $Cert_{GP}$'s private key to obtain user authentication from *HA*, so that only registered *GP*s are able to issue prescriptions.

Verifiability: Digital signature for an anonymous electronic signature AP_P is able to be verified by *HA*'s certificate $Cert_{HA}$'s public key e_{HA}. Since *HA*'s certificate $Cert_{HA}$ has already been posted on a certification authority *CA*'s directory [14], anyone who received AP_P and a prescription *M* is able to verify the content of the anonymous electronic prescription.

Identifiability: The scheme that we propose does not allow any information on the identity of *GP* to be contained in an electronic prescription, so that no one can find who the prescriber was. However, the signer *HA*, who put a digital signature in an anonymous electronic prescription is able to be identified, and this is because *HA* makes a visit to *CA* or a registration agent *RA* to obtain a certificate $Cert_{HA}$ and during the visit the identity of *HA* is revealed.

Non-repudiation: In order for *HA* to create a private key d_{HA} for putting a digital signature in an anonymous electronic prescription, *HA* must cooperate with *CHA* and *LHA*. With no help of *CHA* and *LHA*, *HA* cannot create a key. For this reason, *HA* is not able to deny the legitimacy of an anonymous electronic prescription after its act of putting a digital signature in it.

Prevention of Abuse: d_{HA}, a private key of *HA* that is used to create an anonymous electronic prescription is separately stored from d_{CHA} and d_{LHA} which satisfy $d_{HA}=d_C \cdot d_L \bmod \varphi(N_{HA})$. If the syntax formalism described in an anonymous electronic prescription is not appropriate, *LHA* rejects putting a digital signature. Therefore, it is impossible to use the private key for another purpose.

Confidentiality: Since an anonymous electronic prescription that the patient *P* received is sent only to *PH* who he or she visits to get medicines, no one can know

about the content of the prescription except legitimate *PH*s and the *GP* who wrote the prescription. Also, even *HA*, *CHA* and *LHA*, which are engaged in the act of putting a digital signature in an anonymous electronic prescription, cannot know about the content of the prescription. *HA* is not be able to know about the content of the prescription until *PH* gives the patient medicines and provides the prescription to *HA* for the purpose of receiving health insurance benefits.

Right to Choose a Dispenser: The patient *P* who has an anonymous electronic prescription can select a *PH* freely. Under our scheme, only a dispenser who has been selected by the patient can dispense medicines.

Prevention of Creation of a Cartel: No information on the prescriber is contained in an anonymous electronic prescription. Even if *GP*s try to establish an illegal business tie with pharmacists or drug manufacturers, there's no way that from anonymous electronic prescriptions pharmacists or drug manufacturers can find out which *GP*s have selected their medicines.

Traceability: When a medical dispute has occurred between the patient *P* and the general practitioner *GP* or between the patient *P* and the pharmacist *PH*, an anonymous electronic prescription serves as a crucial clue telling who made a mistake. For this, *CHA* and *LHA* that are engaged in putting a digital signature in the prescription should provide help. First, *LHA* uses a public key e_{HA} to carry out exponential operation for AP_P and calculates $h=(AP_P)^{e_{HA}} \bmod N_{HA}$, and uses a search key h to find TM_5 from its database. After finding TM_5, *LHA* sends it to *CHA*. For TM_5, *CHA* uses *HA*'s public key e_{HA} to carry out exponential operation and calculates $TM_1=(TM_5)^{e_{HA}} \bmod N_{HA}=(TM_2^{d_L})^{e_{HA}} \bmod N_{HA} = ((TM_1^{d_C})^{d_L})^{e_{HA}} \bmod N_{HA}=(TM_1^{d_{HA}})^{e_{HA}} \bmod N_{HA} \bmod N_{HA}$. *CHA* uses a search key TM_1 to find $Cert_{GP}$ in its database. Since $Cert_{GP}$ is the certificate that was issued to *GP* who wrote a prescription, you can find out the identity of *GP* later, if necessary.

Disconnectivity for Doctors: An anonymous electronic description contains no information about *GP*. Except the case that *CHA* and *LHA* cooperate with each other to find out who has signed the electronic prescription, there's no way that anyone can know who the prescriber was--impossible to remove pseudonymity for *GP*, who usually creates multiple prescriptions. As a result, the pharmacy is not able to get any information about *GP* from an electronic prescription.

Prevention of Reuse: Since the information of the state of an anonymous electronic prescription is manages by *HA*, abuse of medicines through the reuse of the same prescription can be prevented. If an anonymous electronic prescription has already been used for dispensing medicines, it is reported to *HA* by *PH*. Therefore, the same prescription cannot be used again. It can be said that *HA* that provides the information of the state of a prescription plays a similar role to that of an OCSP (Online Certificate Status Protocol) server providing the information of the state of a certificate. For this reason, it is impossible for the patient *P* to use a prescription which has already been used to get more medicines[5].

Pseudonymity: An anonymous electronic description does not provide pseudonymity for the patient. To address this problem, if an anonymous certificate is issued to the patient in the same way of issuing an anonymous electronic prescription,

pseudonymity for the patient can be ensured. Pseudonymity for *GP* is ensured because no personal information on *GP* is contained in an anonymous electronic prescription.

Minimum Data Exposure: The content of an anonymous electronic prescription cannot be known by anyone else, except the prescription writer *GP* and the verifier *PH*. The content of an electronic prescription is not revealed until *PH* provides the prescription to *HA* to receive health insurance benefits. Moreover, an electronic prescription is in the patient *P*'s smart card, the description is protected against other people through PIN authentication[6].

5 Conclusion

In this paper, we proposed an RSA-based anonymous electronic prescription scheme, which has no problem with directly applying to RSA-based PKI environments found in most of the countries around the world. This scheme makes sure that anonymity related to an electronic prescription is ensured and also when a medical dispute occurs, the mask of anonymity can be removed in cooperation with the healthcare authority. For these reasons, it is believed that if an electronic ID card policy and a policy proposed by the National Health Insurance Corporation (NHIC) are implemented at the national level, disputes related to the use of electronic prescriptions can be effectively addressed, and transparency in running the national healthcare service program will be improved by preventing the abuse of medicines and potential creation of a cartel between pharmacists and drug manufacturers.

References

1. Lee, B., Kim, H., Kim, K.: Strong proxy signature and its applications. In: Proc. SCIS, pp. 603–608 (2001)
2. Yang, Y., Han, X., Bao, F., Deng, R.H.: A Smart-Card-Enabled Privacy Preserving E-Prescriptions System. IEEE Transactions on Information Technology in Biomedicine 8(1), 47–58 (2004)
3. Medicine prescribing and dispensing support system. Health Insurance Review & Assessment Service, http://www.hira.or.kr/rfl_dur_freeboard_intro_01.do
4. Chadwick, D., Mundy, D.: The secure electronic transfer of prescriptions. In: Healthcare Computing 2004, BCS HIC, pp. 11--25 (2004)
5. Lee, K., Won, D., Kim, S.: A Secure and Efficient E-Will System Based on PKI. Information - An International Interdisciplinary Journal, International Information Institute 14(7), 2187–2206 (2011)
6. Lee, C., Lee, K., Kim, S., Won, D.: Analysis on Vulnerability of Home Healthcare Medical Devices and Development of Protection Profile based on Common Criteria Version 3.1. In: Proc. of CNSI 2011, International Conference on Computers, Networks, Systems, and Industrial Engineering, Jeju Island, May 23-25, pp. 240–247 (2011)

Advanced Malware Variant Detection Algorithm Using Structural Characteristic of Executable File*

Donghwi Shin, Kwangwoo Lee, and Dongho Won**

Information Security Group,
School of Information and Communication Engineering, Sungkyunkwan University,
300 Cheoncheon-dong, Jangan-gu, Suwon, Gyeonggi-do 440-746, Korea
{dhshin,kwlee,dhwon}@security.re.kr

Abstract. The malware is not the strange word. As much again the threat and number of malware is raising. However we and vendors do not arrange these malwares. They store these malware. The reason is the number is so many and the detail analysis is impossible. Therefore some researchers or vendor implemented the behavior based analysis system. However, these systems just analyze malware and do not arrangement. Although there are some arrangement or detection algorithms, there are just results of paper and are not implemented. In this paper, we propose the more available algorithm. The proposed algorithm is the updated version of the CFG (Control Flow Graph) based algorithm.

Keywords: Malware, Control Flow Graph, CFG.

1 Introduction

The malware is not just malicious software. Now that is an industry and there is a market based on the principle of market economics. In other words, some people take an economic benefit. Therefore recently many people create and spread a malware. Since there is a lot of malware, these vendors cannot analyze each malware in detail. Therefore some venders or researcher develops an automatic analysis system. However some malwares include the automatic analysis system detection technique. Therefore in academic world, they are researching an automatic static analysis. The typical sample is the BitBlaze project. Of course, their system cannot analyze the whole malware and detect malware. And there is another issue. That is a relation with between malware. This issue is caused by a malware builder and a generator. And the human decides the relation through the result of analysis from human. Therefore the decision of human may be include an experiment or subjective opinion. In this paper, we propose the algorithm to provide so simple and great analysis performance. In this algorithm, it will attempt to use the control flow graph information for the relation

* 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(2011-0026023).

** Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 301–308, 2011.

analysis. The organization of this paper is as follows: In section 2, we describe some related works. In section 3, we describe the method for extracting CFG (Control Flow Graph) information. In section 4, we describe a similarity calculation of between CFG information and the result of experiment based on a similarity calculation. In section5, we describe another method that determines the relation. Finally, we will conclude the search in section 6.

2 Related Works

In this section we describe the CFG. The CFG is the abbreviation of "Control Flow Graph". The component of CFG is a node and edge. As we know the mean from the name, the node is a basic block and the edge is a path between these nodes. In general, the CFG has a structural characteristic of executable file. Therefore, we use the structural characteristic for a relation analysis in this paper.

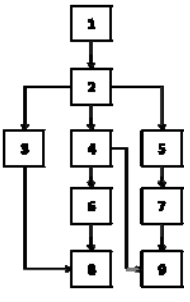


Fig. 1. CFG

Because of the characteristic of CFG, there are some researches using CFG. We will describe some research using CFG. In 2000, Silvio Cesare provided the analysis method using CFG. He calculated the edit distance against an intermediate language and got the similarity. So we think that the main contribution of his research may be a translation according an instruction semantic. However there is not an enough experiment result in his paper. And the Marius Gheorghescu applied three algorithms, edit distance, inverted index, and bloom filters, for a variant identification and compared a performance between these algorithms. He thought a problem of previous researches was a reasonable time for an analysis. He used only basic block from CFG for identification. And he excluded library function calls from CFG basic block. He also masked offset with zero. We thought these processes would reduce variant calculation time. There is another research. The Christopher Kruegel used a structural characteristic of executable file for the polymorphic worm detection and the structural characteristic was extracted from CFG. And the graph coloring technique was used to reflect the instruction semantic. And in his research, there was major contribution. He specified requirements of CFG base signature.

1. Uniqueness
2. Robustness to insertion & deletion
3. Robustness to modification

Except above-mentioned researches, Goldberg, Wehner, Carrera, Karim analyzed malware mutual relation or variant and classified these by a structural characteristic. But, the major problem of these researches is a false positive, applicability, and performance including time complexity.

3 CFG Information Extraction

In this section, we will describe an extraction method as of both the CFG information (basic block and edge) and another information for an instruction semantic. Figure 2 shows the CFG based analysis system diagram.

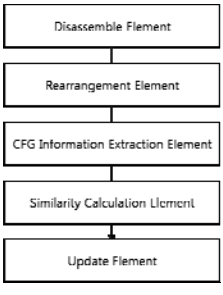


Fig. 2. CFG based Analysis System Diagram

To extract the CFG information, we disassemble a text area of executable file, is not packed or unpacked. However the output of disassemble element is not arranged in a function as a unit. It is not easy to extract basic information. So the rearrangement element divides the output of disassemble element into a function as a unit. The reason of CFG information extraction with a function as a unit is that a malware is modularized and there is some builder. And the CFG information extraction element extracts information from CFG for a similarity calculation. The extracted information is as following.

Number of basic block in function, Number of edge in function
Number of call instruction in function, Number of push instruction in function

The reason why we count the number of call and push instruction is to solve the problem of previous research and to get an advantage of the instruction semantic. And it will satisfy three requirements of signature that the Christopher Kruegel proposed. The both the number of basic block and edge reflects the structural characteristic of CFG. Although node and edge reflects the structural characteristics, these do not reflect an instruction semantic. Therefore to represent an instruction semantic in function, we selected an instruction opcode. Then we have to select the most suitable

instruction to reflect the instruction semantic. At first, we excluded a jump instruction. (ex. jmp, jz, je, and so on) because the jump instruction was used when we extracted a basic block from CFG. And the move instruction as the one of most used instruction also was excluded. Finally, the most considerable instruction will be "call" and "push" instruction. These two instructions have a correlation. We can understand the correlation in function call. Before a function calls another function, the function pushes some arguments into the stack. The number of push instruction may be the number of arguments before function call.

We extracted all information for a similarity calculation. And we represented the extracted information like the coordinate value in fourth dimension.

(p, q, r, s) = (number of basic block, number of edge, number of call instruction, number of push instruction)

Now, we needed to check a theoretical effectiveness about this information. And then we compared the information that was extracted from induc and kido malware sample. In result, the extracted information from induc variants was so similar. And the comparison result with induc and kido malware sample was so different. However, there was values such as [1, 0, 0, 0] or [1, 0, 0, 1]. These did not satisfy uniqueness requirements of the Christopher Kruegel. So we need to a filter and have to select the criteria for a filter. According to our result of experiment, we determine the criteria of filter as followings.

The number of basic block > 2, The number of edge >2

The number of call instruction > 3, The number of push instruction > 5

4 Similarity Calculation

In this section, we will describe a similarity calculation that is a role of Similarity Calculation Element and we will present our experiment result. We represented the extracted information like a point on four dimensional spaces. For example, if there is the same point in three dimensional spaces, the distance between these points is zero. If these points are not the same and so close, the distance is close to zero. In our case, we can calculate a distance between two points in four dimensional space as following figure that is similar to a distance calculation in three dimensional space.

$$d = \sqrt{(p_1 - p_2)^2 + (q_1 - q_2)^2 + (r_1 - r_2)^2 + (s_1 - s_2)^2}$$

Fig. 3. Distance Calculation

Then we can calculate a similarity and collect the distance in set C.

S = {d_i | d_i=distance between x_i and y_i, x_i = CFG information from malware X, y_i = CFG information from malware Y}

However, though two functions are not the same, each element for a calculation may be the same. So we cannot think that the number of zero distance is only a criterion of comparison. Therefore we set a range of distance and calculate a ratio (%) in range.

$$\text{Ratio } R_i \text{ of Area i} = \text{entity number of area i} / n(C) * 100,\ 0 \le i \le 5$$

And we assigned the weight value to each area.

Table 1. Weighted Value

Section	Area	Weighted Value
0	0	50(W_0)
1	1 ~ 5	25(W_1)
2	6 ~ 10	13(W_2)
3	11 ~ 15	6(W_3)
4	16 ~ 20	4(W_4)
5	21 ~ 25	2(W_5)
6	26 ~ 30	0
7	31 ~ 35	0
8	35 ~ 40	0
9	40 ~	0

We assigned only a weighted value to six areas like above table. Finally, we can calculate a similarity like the following algorithm.

$$\text{if } R_i == MAX(R_i) \text{ then } S_i = W_i \text{ else } S_i = 0, \quad 0 \le i \le 5$$

$$\text{Similarity} = S_0 + S_1 + S_2 + S_3 + S_4 + S_5$$

We experiment our algorithm for 177 malware. In this experiment, we could evaluate our algorithm. For example, in case the "Kido" sample, we could identify these variant exactly. In other word, when we calculated the point through our algorithm, a sample of maximum point is "Kido". And the maximum zero matched sample also a "Kido". However the name of closest sample will be able to be equal to the input malware name. So we did think that the result was not absolutely and analyzed a relation by our algorithm. In following diagram, the red line means that it has the highest point and the blue line means that a sample has the most zero distance. Though the name of closest sample is not “Induc”, we are able to draw the following diagram. We get the closest sample of "Induc" is “Agent” by the relation analysis.

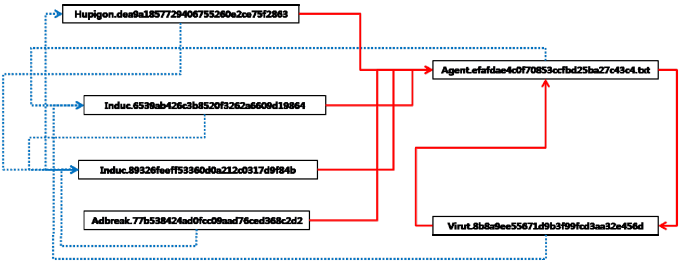


Fig. 4. Induc Relation Diagram

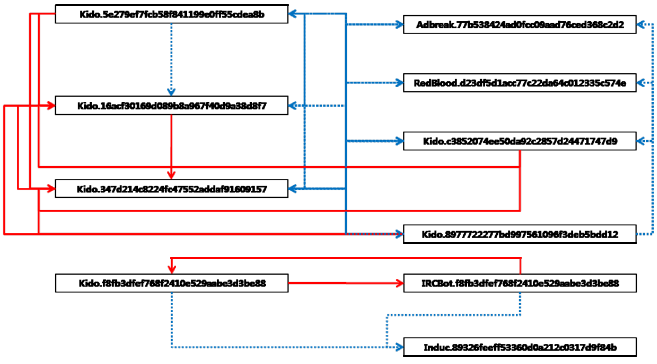


Fig. 5. Kido Relation Diagram

Therefore our algorithm can present the similarity and relation. Through the similarity and relation, we are able to know the variant. From above experiment, we were able to verify that our algorithm identify variant information. The following table is the part of our experimental result.

Table 2. Experimental Result

Input	50	25	13	6	4	2	Max.0
Agent	Bacteraloh	BO2K	Kido	Kido	Sality	Cdudoor	Induc
Agobot	Agobot	BO2K	Delf	AutoSpy	Agent	Cdudoor	Induc
Aimaster	Aimaster	BO2K	Agent	AutoSpy	BO2K	Ataka	Aimaster
BO2K	BO2K	BO2K	Virut	Delf	Cdudoor	Cdudoor	Induc
Hupigon	Agent	BO2K	Virut	AutoSpy	Cdudoor	Cdudoor	Induc
Induc	Agent	BO2K	Virut	AutoSpy	Cdudoor	Cdudoor	Induc
Kido	Kido	Virut	Asylum	Pakes	Barbie	BO2K	Kido
Virut	Virut	BO2K	Agent	Autodpy	Cdudoor	Cdudoor	Virut

From Table 2 and Fig. 4~5, we could identify the relation and variant between samples. However, there may be some mistakes. So we updated our algorithm that considered function sequence. The updated algorithm will be described in section 5. And through updated algorithm, we will completely identify the variant and relation.

5 Function Sequence Comparison

In section 4, we described the method of similarity calculation. The method compared with vis-a-vis functions. As above mentioned, there is the case that even though we calculate a distance with different malware, there may be zero distance. And there was a gray area in some cases. So we needed another comparison factor and we considered the function sequence. If you extract the CFG information, you are able to know that CFGs from similar malwares are so similar and the sequence also is similar.

We defined the "window" as a function sequence length. And we compare function sequences from two malware. In this section, we applied an algorithm with "BO2K", "Zbot", "Agobot", "Aimaster", "Ainder" and "Sality" samples. And the window size is 3. In our experiment, we used two "Agobot" malware samples. Through our algorithm, their sequences were equal to each other as following figure. And these "Agobot.4037a23b1195c3ca54dde79c175e8e4b" malware samples were equal to some malware samples partially. This partial match is derived from the small window size.

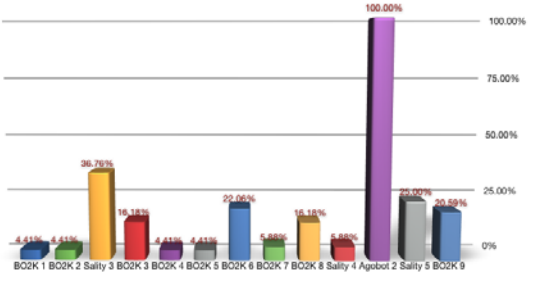


Fig. 6. Abogot.4037a23b1195c3ca54dde79c175e8e4b Comparison Graph

And in "BO2K.089c0352383818e85d6b7219aae26397e" case, there were so many samples that included the same sequence. This result was also derived from a small window size. A close look at the output identifies the variant exactly. In following figure, the input "BO2K.089c0352383818e85d6b7219aae26397e" is matching up the "BO2K4" as a probability 85.83%. And it is match up the "BO2K5" as a probability 63.33% and the "BO2K7" as a probability 64.17%".

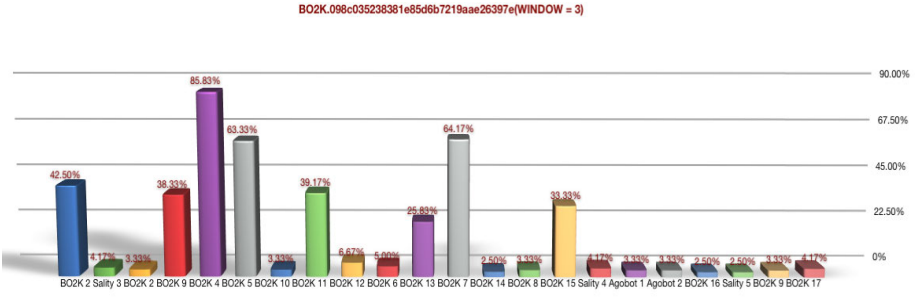


Fig. 7. BO2K.089c0352383818e85d6b7219aae26397e Comparison Graph

Because the sequence comparison considered the function sequence, it made up for weak point of the algorithm in pervious section. Though we experimented with windows value 3 and 8, we need to increase or decrease the window value for an accuracy improvement.

6 Conclusion

Nowadays, the progress of human resource and tool is not keeping up with a speed of market development. Therefore many people are researching about an automatic analysis method. Until now, the research about the behavior-based analysis is so active. A malware author tries to evade an automatic analysis system with various techniques. The behavior based analysis system is limited that what do understand a relation or determine the variant through these systems. To solve this problem, we propose the algorithm that analyze the relation and identify variant of some malware using the structural characteristic of executable file. Then we evaluated our algorithm in experiment. Moreover we added a filter and the window concept to improve a detection or identification performance. Finally, we could detect and identify variant information by our algorithm. We will change our comparison method from string comparison to bit comparison for a performance improvement. Then, we will update a value of filter or window to determine a suitable value.

References

1. BitBlaze: Binary Analysis for Computer Security, http://bitblaze.cs.berkeley.edu/
2. Goldberg, L.A., Goldberg, P.W., Phillips, C.A., Sorkin, G.B.: Constructing computer virus phylogenies. Journal of Algorithms 26(1), 188–208 (1998)
3. Carrera, E., Erdélyi, G.: Digital genome mapping – advanced binary malware analysis. In: Proc. Virus Bull. Int. Conf., pp. 187–197 (September 2004)
4. Karim, E., Walenstein, A., Lakhotia, A., Parida, L.: Malware phylogeny using maximal p-patterns. In: Proceedings of the EICAR 2005 Conference, pp. 167–174 (April–May 2005)
5. Gheorghescu, M.: An Automated Virus Classification System. In: Virus Bulletin Conference (2005)
6. Cesare, S., Xiang, Y.: Classification of Malware Using Structured Control Flow. In: Proc. 8th Australasian Symposium on Parallel and Distributed Computing (2010)
7. Krügel, C., Kirda, E., Mutz, D., Robertson, W., Vigna, G.: Polymorphic Worm Detection Using Structural Information of Executables. In: Valdes, A., Zamboni, D. (eds.) RAID 2005. LNCS, vol. 3858, pp. 207–226. Springer, Heidelberg (2006)
8. Zubair Shafiq, M.: PE-probe: leveraging packer detection and structural information to detect malicious portable executables. In: VB 2009 (2009)
9. Kaczmarek, M.: Architecture of a Morphological Malware Detector. Journal in Computer Virology (2008)
10. Vinod, P.: Static CFG analyzer for metamorphic. In: PIN 2009 (2009)
11. Dullien, T.: Rolf Rolles, and Ruhr-universitaet Bochum, Graph-based comparison of executable objects (2005)
12. Sabin, T.: Comparing binaries with graph isomorphisms (2000)

Cryptanalysis of a Group Key Transfer Protocol Based on Secret Sharing*

Junghyun Nam[1], Moonseong Kim[2], Juryon Paik[3], Woongryul Jeon[3], Byunghee Lee[3], and Dongho Won[3,**]

[1] Department of Computer Engineering, Konkuk University, Korea
jhnam@kku.ac.kr
[2] Information and Communications Examination Bureau, Korean Intellectual Property Office, Korea
moonseong@kipo.go.kr
[3] Department of Computer Engineering, Sungkyunkwan University, Korea
wise96@ece.skku.ac.kr,{wrjeon,bhlee,dhwon}@security.re.kr

Abstract. Group key establishment protocols allow a set of communicating parties to establish a common secret key. Due to their significance in building a secure multicast channel, a number of group key establishment protocols have been suggested over the years for a variety of settings. Among the many protocols is Harn and Lin's group key transfer protocol based on Shamir's secret sharing. This group key transfer protocol was designed to work in the setting where a trusted key generation center shares a long-term secret with each of its registered users. As for security, Harn and Lin claim that their protocol prevents the long-term secret of each user from being disclosed to other users. But, we found this claim is not true. Unlike the claim, Harn and Lin's protocol cannot protect users' long-term secrets against a malicious user. We here report this security problem with the protocol and show how to address it.

Keywords: Security, key establishment protocol, group key transfer, secret sharing, replay attack.

1 Introduction

Key establishment protocols allow two or more communicating parties to establish their common secret key called a *session key*. Establishment of session keys is one of the fundamental cryptographic operations and provides a typical way of building secure communication channels over insecure public networks. Traditionally, protocols which can be run by an arbitrary number of parties are called group (or conference) key establishment protocols, in contrast to protocols

* This work was supported by Priority Research Centers Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2011-0018397).

** Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 309–315, 2011.

which can be run only by two or three parties. In the group setting, a session key is also called a *group key*. Key establishment protocols are often classified into two types: key agreement protocols and key transfer protocols. Key agreement protocols require each participant to contribute its part to the final form of the session key, whereas key transfer protocols allow one trusted entity to generate the session key and then transfer it to all participants.

The first priority in designing a key establishment protocol is placed on ensuring the security of the protocol. Even if it is computationally infeasible to break the cryptographic algorithms used, the whole system becomes vulnerable to all manner of attacks if the keys are not securely established. But the experience shows that the design of secure key establishment protocols is notoriously difficult. Over the last decades, a number of protocols have been found to be insecure years after they were published [5,4,2,1]. Thus, key establishment protocols must be subjected to a thorough scrutiny before they can be deployed into a public network which might be controlled by an adversary.

This work is concerned with the security of the group key transfer protocol designed recently by Harn and Lin [3]. We use HL to refer to Harn and Lin's protocol. The protocol HL employs Shamir's secret sharing [6] to achieve information-theoretically secure distribution of session keys. Accordingly, the security of HL does not depend on any unproven assumption about computational hardness (as far as confidentiality of session keys is concerned). HL assumes a trusted key generation center (KGC) who provides key distribution service to its registered users. During registration, KGC issues each user a long-term secret which should be kept privately by the user. One of the security claims made for HL is that the long-term secret of each user cannot be learned by other users. But, it turns out that this claim is not true. The truth is that HL is vulnerable to a replay attack whereby a malicious user, who is registered with KGC, can readily obtain the long-term secret of any other registered user. In the current work, we reveal this security vulnerability of HL and then suggest a countermeasure against the replay attack.

2 Harn and Lin's Group Key Transfer Protocol

This section reviews Harn and Lin's group key transfer protocol HL [3]. The protocol HL consists of three phases: system initialization, user registration, and key distribution.

System Initialization. KGC randomly chooses two safe primes p and q (i.e., p and q are primes such that $p' = \frac{p-1}{2}$ and $q' = \frac{q-1}{2}$ are also primes) and computes $n = pq$. n is made publicly known.

User Registration. Each user is required to register at KGC to subscribe the key distribution service. During registration, KGC shares a secret (x_i, y_i) with each user U_i where $x_i, y_i \in \mathbb{Z}_n^*$.

Key Distribution. This phase constitutes the core of the protocol and is performed whenever a group of users $U_1, \ldots, U_t$ decide to establish a common session key.

Step 1. A designated user of the group, called the initiator, sends a key distribution request to KGC. The request carries the list of participating users $\langle U_1, \ldots, U_t \rangle$.

Step 2. KGC broadcasts the participant list $\langle U_1, \ldots, U_t \rangle$ in response to the request.

Step 3. Each user U_i, for $i = 1, \ldots, t$, sends a random challenge $r_i \in \mathbb{Z}_n^*$ to KGC.

Step 4. KGC randomly selects a session key k and constructs by interpolation a t-th degree polynomial $f(x)$ passing through the $(t+1)$ points: $(x_1, y_1 \oplus r_1), \ldots, (x_t, y_t \oplus r_t)$ and $(0, k)$. Next, KGC selects t additional points $P_1, \ldots, P_t$ that lie on the polynomial $f(x)$. KGC then computes $\beta = h(k, U_1, \ldots, U_t, r_1, \ldots, r_t, P_1, \ldots, P_t)$, where h is a one-way hash function, and broadcasts $\langle \beta, r_1, \ldots, r_t, P_1, \ldots, P_t \rangle$ to the users. All computations with respect to $f(x)$ are performed modulo n.

Step 5. Each U_i constructs the polynomial $f(x)$ from the $(t+1)$ points: $P_1, \ldots, P_t$ and $(x_i, y_i \oplus r_i)$. Then U_i recovers the session key $k = f(0)$ and checks the correctness of β in the straightforward way. U_i aborts if the check fails.

Since the above protocol HL focuses on protecting the keying material broadcasted from KGC to users, Harn and Lin also present (in Remark 2 of [3]) how HL can be extended to provide user authentication and key confirmation. Let HL^+ be the extended version of HL. HL^+ is constructed from HL by revising Steps 3 and 4 to achieve user authentication and by adding Steps 6 and 7 to achieve key confirmation.

Step 3 (of HL^+). Each user U_i, for $i = 1, \ldots, t$, selects a random challenge $r_i \in \mathbb{Z}_n^*$, computes $\alpha_i = h(x_i, y_i, r_i)$, and sends $\langle \alpha_i, r_i \rangle$ to KGC.

Step 4 (of HL^+). KGC checks the correctness of each α_i in the straightforward way. KGC aborts if any of the checks fails. Otherwise, KGC continues with Step 4 of HL.

Step 6. Each U_i sends $\gamma_i = h(x_i, y_i, k)$ to KGC.

Step 7. After receiving all γ_i's, KGC sends $\delta_i = h(x_i, y_i, k, U_1, \ldots, U_t)$ to U_i for $i = 1, \ldots, t$.

All other parts (including the phases of system initialization and user registration) remain unchanged between HL and HL^+.

3 Replay Attack

The fundamental security goal of a key establishment protocol is to ensure that no one other than the intended users can compute the session key. In the cases of HL and HL^+, this goal can be achieved only when the secrecy of every (x_i, y_i) is guaranteed. As soon as (x_i, y_i) is disclosed, all the protocol sessions that U_i participates become completely insecure. It is thus crucial that x_i's and y_i's must not be revealed under any circumstances.

Harn and Lin claim that their protocols prevent the secret (x_i, y_i) of each U_i from being disclosed to other users, either insiders or outsiders (Theorem 3

of [3]). However, we found that this claim is wrong. Suppose that a malicious registered user U_j has a goal of finding out U_i's secret (x_i, y_i). Then U_j can achieve its goal by mounting the following attack against the protocol HL^+.

Step 0. As a preliminary step, the adversary U_j eavesdrops on a protocol session, where U_i participates, and stores the message $\langle \alpha_i, r_i \rangle$ sent by U_i in Step 3 of the session.

U_j then initiates two concurrent sessions S and S′ of the protocol alleging that the participants of both sessions are U_i and U_j. Once KGC responds with the participant list $\langle U_i, U_j \rangle$ in Step 2 of each session, U_j performs Step 3 of the sessions while playing dual roles of U_j itself and the victim U_i.

Step 3 of S. U_j sends the eavesdropped message $\langle \alpha_i, r_i \rangle$ to KGC as if the message is from U_i. But, U_j behaves honestly in sending its own message; U_j selects a random $r_j \in \mathbb{Z}_n^*$, computes $\alpha_j = h(x_j, y_j, r_j)$, and sends $\langle \alpha_j, r_j \rangle$ to KGC.

Step 3 of S′. U_j replays the messages $\langle \alpha_i, r_i \rangle$ and $\langle \alpha_j, r_j \rangle$. That is, U_j sends $\langle \alpha_i, r_i \rangle$ as U_i's message and sends $\langle \alpha_j, r_j \rangle$ as its own message.

KGC cannot detect any discrepancy since α_i and α_j are both valid. Note that KGC does not check for message replays. Hence, KGC will distribute the keying materials for the sessions. Let $f(x) = a_2x^2 + a_1x + k$ and $f'(x) = a'_2x^2 + a'_1x + k'$ be the polynomials constructed by KGC respectively in sessions S and S'. As soon as receiving the keying materials, U_j derives these polynomials as specified in Step 5 of the protocol. Now let

$$\begin{aligned} g(x) &= f(x) - f'(x) \\ &= (a_2 - a'_2)x^2 + (a_1 - a'_1)x + k - k'. \end{aligned}$$

Then, $g(x_i) = 0$ and $g(x_j) = 0$ since $f(x_i) = f'(x_i) = y_i \oplus r_i$ and $f(x_j) = f'(x_j) = y_j \oplus r_j$. This implies that x_i and x_j are the two roots of the quadratic equation $(a_2 - a'_2)x^2 + (a_1 - a'_1)x + k - k' = 0$. It follows that

$$(a_2 - a'_2)x^2 + (a_1 - a'_1)x + k - k' = (a_2 - a'_2)(x - x_i)(x - x_j)..$$

Therefore,

$$x_i = x_j^{-1}(a_2 - a'_2)^{-1}(k - k'). \tag{1}$$

Here, the computations are done modulo n. Once x_i is obtained as in Eq.. (1), y_i can be easily computed from $f(x_i) = y_i \oplus r_i$. The value of y_i is different depending on whether $y_i \oplus r_i < n$ or $y_i \oplus r_i \geq n$.

$$y_i = \begin{cases} f(x_i) \oplus r_i & \text{if } y_i \oplus r_i < n \\ (f(x_i) + n) \oplus r_i & \text{otherwise.} \end{cases}$$

α_i can serve as a verifier for checking which of the two cases is true. Using (x_i, y_i) obtained as above, U_j is able to complete the protocol without the attack being noticed.

The above attack assumes, for ease of exposition, that KGC allows for the key establishment between two parties. But, this assumption is not necessary.. If two-party key establishments are not allowed, U_j can collude with another malicious user U_k to mount a slight variant of the attack. Assume two concurrent sessions of the protocol, in both of which the participants are U_i, U_j and U_k. If U_j and U_k collude together and run the two sessions as in the attack above, they can construct a cubic polynomial $g(x) = (a_3 - a'_3)x^3 + (a_2 - a'_2)x^2 + (a_1 - a'_1)x + k - k'$ such that $g(x_i) = g(x_j) = g(x_k) = 0$. Then, with x_j and x_k in hand, the adversaries can compute x_i as

$$x_i = (-1)x_j^{-1}x_k^{-1}(a_3 - a'_3)^{-1}(k - k')$$

and thereby can determine y_i as above.

So far, we have seen the vulnerability of the protocol HL^+. As can be expected, the basic protocol HL also suffers from the same vulnerability. The attack against HL is essentially similar to the above attack and is provided in Appendix.

4 Countermeasure

The security failure of HL^+ (and HL) is attributed to one obvious flaw in the protocol design: the messages sent by users in Step 3 can be replayed in different protocol sessions. This flaw allows our adversary U_j to send the same random challenges twice and thereby to construct a quadratic polynomial $g(x)$ such that $g(x_i) = g(x_j) = 0$.. Fortunately, message replays can be effectively prevented if Steps 2 and 3 of the protocols are revised as follows:

Step 2 (revision). KGC selects a random $r_0 \in \mathbb{Z}_n^*$ and broadcasts it along with the participant list $\langle U_1, \ldots, U_t \rangle$.

Step 3 (revision). Each user U_i, for $i = 1, \ldots, t$, selects a random $r_i \in \mathbb{Z}_n^*$, computes $\alpha_i = h(x_i, y_i, r_i, r_0, U_1, \ldots, U_t)$, and sends $\langle \alpha_i, r_i \rangle$ to KGC..

The other steps of the protocols remain unchanged except that in Step 4 of HL, KGC has to check the correctness of α_i, for $i = 1, \ldots, t$, before starting to construct the polynomial $f(x)$. As a result of our modification, the protocols HL and HL^+ become identical except that HL^+ requires two additional steps (Steps 6 and 7) for key confirmation. With the modification applied, the message $\langle \alpha_i, r_i \rangle$ eavesdropped in a protocol session can no longer be replayed in any other sessions. Hence, our attacks are not valid against the improved protocols.

5 Concluding Remark

It is worth noting that polynomial interpolation over $\mathbb{Z}_n^*$ may fail, though the probability of failure is negligible. This is essentially because the multiplicative group $\mathbb{Z}_n^*$ is not closed under addition and subtraction while interpolation formulas include additive and subtractive terms. If an addition/subtraction operation in $\mathbb{Z}_n^*$ returns a value z such that $\gcd(z, n) \neq 1$ (i.e., $z = cp$ or $z = cq$ for

some integer c), then there will not exist the multiplicative inverse of z modulo n. The protocols HL and HL$^+$ fail if such a z happens to be a divisor in the interpolation formula. (Of course, the probability of this occurring should be negligible, because otherwise we have a polynomial-time factoring algorithm.) This correctness issue of the protocols can be addressed simply by replacing $\mathbb{Z}_n^*$ with a prime field $\mathbb{F}_p$ in which interpolation never fails. We believe that the replacement causes no security degradation.

References

1. Choo, K.-K.: Refuting the security claims of Mathuria and Jain (2005) key agreement protocols. International Journal of Network Security 7(1), 15–23 (2008)
2. Choo, K.-K.R., Boyd, C., Hitchcock, Y.: Errors in Computational Complexity Proofs for Protocols. In: Roy, B. (ed.) ASIACRYPT 2005. LNCS, vol. 3788, pp. 624–643. Springer, Heidelberg (2005)
3. Harn, L., Lin, C.: Authenticated group key transfer protocol based on secret sharing. IEEE Transactions on Computers 59(6), 842–846 (2010)
4. Krawczyk, H.: HMQV: a High-Performance Secure Diffie-Hellman Protocol. In: Shoup, V. (ed.) CRYPTO 2005. LNCS, vol. 3621, pp. 546–566. Springer, Heidelberg (2005)
5. Pereira, O., Quisquater, J.-J.: A security analysis of the Cliques protocols suites. In: Proc. 14th IEEE Computer Security Foundations Workshop, pp. 73–81 (2001)
6. Shamir, A.: How to share a secret. Communications of the ACM 22(11), 612–613 (1979)

Appendix: Attack on the Protocol HL

Consider the protocol HL described in Section 2. We here show that HL is vulnerable to an attack whereby a registered user U_j can learn the long-term secret (x_i, y_i) of any other registered user U_i. The main idea of the attack is the same as that of the attack against HL$^+$. The attack works as follows:

1. The adversary U_j initiates two concurrent sessions S and S$'$ of the protocol alleging that the participants of both sessions are U_i and U_j.
2. KGC will respond with the participant list $\langle U_i, U_j \rangle$ in Step 2 of each session.
3. U_j performs Step 3 of the sessions while playing dual roles of U_j itself and the victim U_i.

   Step 3 of S. U_j selects a random $r_i \in \mathbb{Z}_n^*$ and sends it to KGC as if it is from U_i. In addition, U_j sends its own challenge $r_j \in \mathbb{Z}_n^*$ to KGC.

   Step 3 of S$'$. U_j replays the challenges r_i and r_j. That is, U_j sends r_i as U_i's challenge and sends r_j as its own challenge.
4. KGC will distribute the keying materials for the sessions. Let $f(x) = a_2x^2 + a_1x + k$ and $f'(x) = a_2'x^2 + a_1'x + k'$ be the polynomials constructed by KGC respectively in S and S'.

5. After receiving the keying materials, U_j recovers the two polynomials $f(x)$ and $f'(x)$ as specified in Step 5 of the protocol. Let

$$\begin{aligned} g(x) &= f(x) - f'(x) \\ &= (a_2 - a'_2)x^2 + (a_1 - a'_1)x + k - k'. \end{aligned}$$

Then since $g(x_i) = 0$ and $g(x_j) = 0$,

$$x_i = x_j^{-1}(a_2 - a'_2)^{-1}(k - k').$$

Given x_i, we should consider two different cases in calculating the value of y_i.

$$y_i = \begin{cases} f(x_i) \oplus r_i & \text{if } y_i \oplus r_i < n \\ (f(x_i) + n) \oplus r_i & \text{otherwise.} \end{cases}$$

6. U_j needs to decide whether $y_i \oplus r_i < n$ or $y_i \oplus r_i \geq n$. Note that the equations $f(x_i) = y_i \oplus r_i$ and $f'(x_i) = y_i \oplus r_i$ do not allow to determine which of the two cases is true. This is because both values of y_i satisfy the equations. But, knowledge of another polynomial $f''(x)$ such that $f''(x_i) = y_i \oplus \tilde{r}_i$, where $\tilde{r}_i \neq r_i$, would immediately reveal which one of the two values of y_i is correct. U_j can easily generate such a polynomial $f''(x)$ if he runs a new protocol session as above, but this time using a different challenge $\tilde{r}_i$ for U_i.

As is the case for HL^+, HL is also vulnerable to a colluding attack where two adversaries U_j and U_k collude together to learn U_i's secret (x_i, y_i). We here omit the description of the colluding attack on HL since it is clear from the attack above and the colluding attack on HL^+.

Protection Profile for Data Leakage Protection System*

Hyun-Jung Lee and Dongho Won**

Information Security Group,
School of Information and Communication Engineering,
Sungkyunkwan University, 300 Cheoncheon-dong, Jangan-gu,
Suwon, Gyeonggi-do 440-746, Korea
{hjlee,dhwon}@security.re.kr

Abstract. Most of the biggest financial or insurance companies in the world that deal with critical client information are now considering introducing a Data Leakage Protection(DLP) system in order to reduce the risk of data loss. However, there is no standard for the introduction of a DLP system, which leads to companies still having the problem of information leakage even after the introduction. This paper analyzes various channels of information leakage and suggests a protection profile based on the CC V3.1 to help develop and introduce a DLP system that can prevent in-house information leakage.

Keywords: DLP, Data Leakage Protection, Data Loss Prevention, Protection Profile, Common Criteria.

1 Introduction

Recent reports show that the number of the cases of personal information stolen or lost is increasing and the scale of damage is growing. The IT Policy Compliance Group says 20% of companies suffer from more than 22 cases of information theft a year. According to the Forrester Research, one of the most famous market analysis companies, direct loss resulting from data leakage costs a non-financial company $15 per each client, which includes expenses for notifying clients, credit monitoring services, IT restoration, decline in sales because of past customers, and legal procedures and audit.

The prime reason for in-house information leakage is the lack of security awareness of users. It is not easy to control the users, which are fluent as anyone can assume. The best a company can do is to make sure it has a security system infrastructure for internal security. That is why more and more companies are using a DLP solution, which is a kind of document leakage protection solution, to encode documents and minimize damage in case of leakage. This solution monitors and controls possible channels of information leakage like an email or messenger and traces information leakage due to an individual's intended illegal action. It can

* 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(2011-0026023).

** Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 316–326, 2011.

combine existing solutions and provide more extensive security processes in one step. "Information Protection and Control" report of the IDC in 2007 estimated that the size of DLP market all over the world would be at 3.2 billion dollars by 2011[19].

A DLP solution has been first introduced in financial and insurance industries with a firm standard in handling clients' information, which is understandable considering how it deals with a large amount of major client information. It is now being used in most banks and insurance companies in the U.S. As information is considered a key to a successful business, there are increasing interest and needs for a DLP in many other industries. For example, a manufacturing or high-tech company should protect its information to win the competition; a medical institution should protect the information of patients; and a government institution should protect the information of people and also the confidential information of national defense. Even though the need for a DLP system keeps increasing, there is no standard regarding the security functions required of a DLP system yet, which may lead to a company using a DLP system and still risking information leakage. This paper intends to derive necessary security functions of a DLP system based on the Common Criteria and suggest a baseline of a system that can prevent information leakage.

2 Overview of DLP System

2.1 Definition

Data Loss Prevention (DLP) is a computer security term referring to systems that identify, monitor, and protect data in use (e.g. endpoint actions), data in motion (e.g. network actions), and data at rest (e.g. data storage) through deep content inspection, contextual security analysis of transaction (attributes of originator, data object, medium, timing, recipient/destination and so on) and with a centralized management framework. Systems are designed to detect and prevent unauthorized use and transmission of confidential. Thus the Key Defining Characteristics are "Deep Content Analysis", "Central Policy management" and "Broad Content Coverage across multiple platforms and locations" [5, 8].

2.2 DLP Type

Network DLP. Typically a software or hardware solution that is installed at network egress points near the perimeter. It analyzes network traffic to detect sensitive data that is being sent in violation of information security policies. The DLP Network product detects sensitive data while it is being transmitted across the network, and generates events and incidents reflecting policy violations. The targeted data is referred to as "Data In Motion[21]". DLP Network can automatically monitor or block identified transmissions, or quarantine messages that may need prior approval before leaving the network. In addition, encryption of emails containing sensitive content can be performed by the operational environment when the TOE is configured to do so[18].

End-Point DLP. Such systems run on end-user workstations or servers in the organization. Like network-based systems, endpoint-based can address internal as well as external communications, and can therefore be used to control information flow

between groups or types of users (e.g. 'Chinese walls')[21]. They can also control email and Instant Messaging communications before they are stored in the corporate archive, such that a blocked communication (i.e., one that was never sent, and therefore not subject to retention rules) will not be identified in a subsequent legal discovery situation. Endpoint systems have the advantage that they can monitor and control access to physical devices (such as mobile devices with data storage capabilities) and in some cases can access information before it has been encrypted. Some endpoint-based systems can also provide application controls to block attempted transmissions of confidential information, and provide immediate feedback to the user. They have the disadvantage that they need to be installed on every workstation in the network, cannot be used on mobile devices (e.g., cell phones and PDAs) or where they cannot be practically installed (for example on a workstation in an internet café).

Storage DLP. Typically a software solution that is installed in data centers to discover confidential data is stored in inappropriate and/or unsecured locations (e.g. open file share)[21].

3 Proposed Data Leakage Protection System Protection Profile

In this section, we describe the proposed DLP system and protection profile.

3.1 Overview of TOE

The most ideal way of data leakage protection is to prevent a leakage from happening in the first place by understanding what needs to be protected, identifying every path or channel that could be used for stealing data, and monitoring them. Since one measure alone cannot block all paths, it seems to be the best way to combine different types of DLPs. This paper suggests an environment where End-Point DLP, Storage DLP, and Network DLP can be operated together and provide a protection profile (PP) to which all these types can conform.

3.2 Security Problem Definition

This section describes the security aspects of the environment in which the TOE will be used and the manner in which the TOE is expected to be employed. It provides the statement of the TOE security environment, which identifies and explains all:

- Known and presumed threats countered by either the TOE or by the security environment
- Organizational security policies with which the TOE must comply
- Assumptions about the secure usage of the TOE, including physical, personnel and connectivity aspects

Threats. This subsection of the security problem definition shows the threats that are to be countered by the TOE. A threat consist of a threat agent, an asset and an adverse action of that threat agent on that asset[2]. The specification of threats should include

all threats detected up to now, if it is not done the TOE may provide inadequate protection. In other words, if the specification of threats is insufficiency, the assets may be exposed to an unacceptable level of risk. In the result, we derive the threats in table 1[6,7,8,11,12,13,14,15,16,17].

The threat agents are divided into two categories:

- Attackers who are not TOE users: They have public knowledge of how the TOE operates and are assumed to possess a low skill level, limited resources to alter TOE configuration settings or parameters and no physical access to the TOE.

Table 1. Threats

Name	***Description***
T.Spoofing	A threat agent can access the TOE by disguising himself as an authorized user.
T.Consecutive Authtication Attempt	A threat agent can attempt authentication consecutively and obtain the administrator's privilege.
T.Storaed TSF Data	An unauthorized user may try accessing the TOE or attack the system in order to change or delete the TOE configuration or get access to the functionality or data of the TOE.
T.TSF Data Transfer	The TSF data can be exposed or modified while it is being transferred between different components of the TOE or between the TOE and a remote administrator.
T.Recored Failure	A threat agent can cause an action that will exhaust the storage so that the TOE fails to record security-relevant events.
T.Reuse Attack	A threat agent can reuse the authentication data of an authorized user to access the TOE.
T.Unauthorized Process	A threat agent can start an unauthorized process on a user's PC to steal the user data or stop the TOE.
T.User Data Leakage	A threat agent can leak the user data without authorization.
T.Failure (End-Poing DLP Only)	A threat agent can take advantage of the TOE_Agent when it is not capable of providing services because an attack caused failure.
T.Mnagement Failure	The TOE may fail to take care of a threat agent's inappropriate access to or action taken on the data that needs protection, which may result in the data modified or tampered with.
T.Analysis Failure	The TOE may fail to detect a threat agent's inappropriate access to or action taken on the data that needs protection, which may result in the data modified or tampered with.
T.Unaythorized Action	A threat agent can access the data that needs to be protected and take unauthorized actions on it.
T.Unauth	A user can access the TOE and its data even when he is not authorized according to the security policy of the TOE.
T.Compromise	An unauthorized user can bypass the security mechanism to expose, eliminate, or destroy the integrity of the data collected or generated by the TOE.
T.Application Execute	An attacker can start inappropriate software on the system or make inappropriate changes to the system without being caught.

- TOE users: They have extensive knowledge of how the TOE operates and are assumed to possess a high skill level, moderate resources to alter TOE configuration settings or parameters and physical access to the TOE. (TOE users are, however, assumed not to be willfully hostile to the TOE.)

Both are assumed to have a low level of motivation. The IT assets requiring protection are the user data saved on or transitioning through the TOE and the hosts on the protected network.

Organizational Security Policy. An Organizational Security Policy (OSP) is a set of security rules, procedures, or guidelines imposed by an organization on the operational environment of the TOE. The organizational security policies for this paper are described in Table 2.

Table 2. Organizational Security Policy

Name	*Description*
P.Audit	The TOE shall generate and maintain a record of security-related events to ensure accountability. Records shall be reviewed.
P.Secure Management	The TOE shall provide its authorized administrator with a means to manage the TOE securely and keep the TSF data up to date.
P.Statistics	An authorized administrator shall be able to take statistics on the data of audit and intrusion detection.

Assumptions. The assumptions are made on the operational environment in order to be able to provide security functionality. If the TOE is placed in an operational environment that does not meet these assumptions, the TOE may not be able to provide all of its security functionality anymore. Assumptions can be on physical, personnel and connectivity of the operational environment.

Table 3. Assumptions

Name	*Description*
A.Dynamic Managment	The TOE is managed in a way that it can deal with dynamic changes of the assets that need to be protected.
A.Physical Security	The TOE locates in a secure environment that only an authorized administrator can access. (Except for End-point DLP Agent)
A.Trusted Admin	An authorized administrator of the TOE has no malicious intention, is properly educated in terms of the management functions of the TOE, and follows the administrator's guidance.
A.OS Reinforcement	The TOE has a routine to remove unnecessary services or measures and to fix vulnerability of the OS(e.g. using patches) to ensure credibility and stability of the OS.
A.Access	The TOE can access all IT system data required to enforce its functionality.
A.Timestamp	The IT environment provides the TOE with a reliable timestamp.

3.3 Security Objectives

Security objectives are concise, abstract statements of the intended solution to the problem defined by the security problem definition. The set of security objectives for a TOE form a high-level solution to the security problem. This high-level solution is divided into two part-wise solutions: the security objectives for the TOE, and the security objectives for the TOE's operational environment. This section identifies the security objectives for the TOE and its supporting environment.

Security Objectives for the TOE. The TOE provides security functionality to solve a certain part of the problem defined by the security problem definition. This part wise solution is called the security objectives for the TOE and consists of a set of objectives that the TOE should achieve in order to solve its part of the problem[3].

Table 4. Security Objectives for the TOE

Name	*Description*
O.Audit	The TOE shall generate and maintain a record of security-related events to ensure accountability. It shall also provide a means to review the records.
O.MNG	The TOE shall provide its authorized administrator with an efficient means to manage the TOE and keep the TSF data up to date.
O.Data Protection	The TOE shall protect the TSF data stored in it from unauthorized exposure, modification, or deletion.
O.IA	The TOE shall uniquely identify a user and authenticate the user before allowing his access to the TOE.
O.Security Access	Only authorized administrator shall be allowed to access the security functionality, configuration, and data.
O.Data Collection	The TOE shall collect from the managed system the program codes that can be allowed and objects that need to be protected.
O.Data Analysis	The TOE shall have an analysis process to decide whether to allow or deny access of an object.
O.Tagging	The TOE shall be able to identify the data categorized by data analysis.
O.Leakage Protection	The TOE shall monitor itself to prevent leakage of assets.
O.Audit Review	The TOE shall provide the authorized administrator with functionality to filter, review, and order the audit records.
O.Statistics	The TOE shall analyze and take statistics on all events according to the policy.
O.Notice	The TOE shall raise an alarm according to the policy set for each event.
O.Leakage Management	The TOE shall enforce the policy and take actions on the files that hold confidential or secret information; and on the transfer of, user's action on, or access to the information.
O.self Protection	The TOE shall protect itself from unauthorized access or tampering to its functionality and data in order to maintain the integrity of the system data and audit records.

Operational Environment of the TOE. The operational environment of the TOE implements technical and procedural measures to assist the TOE in correctly providing its security functionality (which is defined by the security objectives for the TOE). This part wise solution is called the security objectives for the operational environment and consists of a set of statements describing the goals that the operational environment should achieve[3].

Table 5. Operational environment of the TOE

Name	*Description*
OE.Dynamic MNG	The TOE shall be managed in a way that it can deal with dynamic changes of the system that needs to be protected.
OE.Physical Security	The TOE shall be located in a secure environment that only an authorized administrator can access.
OE.Trusted Admin	An authorized administrator of the TOE shall have no malicious intention, be properly educated in terms of the management functions of the TOE, and follow the administrator's guidance.
OE.OS Reinforcement	The TOE shall have a routine to remove unnecessary services or measures and to fix vulnerability of the OS(e.g. using patches) to ensure credibility and stability of the OS.
OE.Timestamp	The TOE shall record security-relevant events accurately using the reliable timestamp provided in the operational environment.
OE.Access	The TOE shall be able to access all IT system data required to enforce its functionality.
OE.Timestamp	The IT environment shall provide the TOE with a reliable timestamp.

3.4 Extended Components Definition (ASE_ECD)

This section specifies the extended SFRs for the TOE. The extended SFRs are organized by class. Table 6 identifies all extended SFRs implemented by the TOE and a detailed description of each component is in Table 7.

Table 6. Extended TOE Security Functional Requirements

Class Name	*Component Name*	*Description*
Security Management	EXT_FMT_STA.1	Data statistics of audit and leakage detection
User Data Protection	EXT_FDP_COL.1	Monitored data collection
	EXT_FDP_ANL.1	Monitored data analysis
	EXT_FDP_MON.1	Real-time monitoring of data leakage
	EXT_ FDP_PRV.1	Prevention of data leakage

Table 7. Datailed Description of Extended TOE SFR

EXT_FMT_STA.1	
Family Behavior	Data statistics of audit and leakage detection (FMT_STA, Statistics) This family provides the functionality to take statistics on the data generated as a result of audit and detection of leakage.
Component description & structure	EXT_FMT_STA.1 Data statistics of audit and leakage detection provides an authorized administrator with the capability to take statistics on the data.
	EXT_FMT_STA.1 Data statistics of audit and leakage detection Hierarchical to: No other components Dependencies: FAU_GEN.1 Audit data generation EXT_FDP_ANL.1 Monitored data analysis EXT_FDP_REV.1 Prevention of data leakage ***EXT_FMT_STA.1.1 The TSF shall provide [assignment: an authorized user] with the capability to take statistics on the data generated as a result of audit and detection of leakage.***
EXT_FDP_COL.1	
Family Behavior	Monitored data collection (EXT_FDP_COL, Collection) This family intends to collect the data for monitoring from the user data in the protected system.
Component description & structure	EXT_FDP_COL.1 Monitored data collection collects the data for monitoring from the protected system to prevent data leakage.
	EXT_FDP_COL.1 Monitored data collection Hierarchical to: No other components Dependencies: No dependencies ***EXT_FDP_COL.1.1 The TSF shall collect [assignment: list of data that needs to be collected] from the protected system according to [assignment: criteria of data collection] to prevent data leakage.*** ***EXT_FDP_COL.1.2 The data collected by the TSF shall include the following information:*** ***a) Date and time of the event*** ***b) [Assignment: additional attributes of the data for monitoring]***
EXT_FDP_ANL.1	
Family Behavior	Monitored data analysis (EXT_FDP_ANL, Analysis) This family intends to analyze and identify the data collected for monitoring.
Component description & structure	EXT_FDP_ANL.1 Monitored data analysis analyzes and identifies the collected data to prevent data leakage.
	EXT_FDP_ANL.1 Monitored data analysis Hierarchical to: No other components Dependencies: EXT_FDP_COL.1 Monitored data collection ***EXT_FDP_ANL.1.1 The TSF enforces the following analysis functionality based on the collected data.*** ***a) [Assignment: function to analyze the data]*** ***EXT_FDP_ANL.1.2 The TSF shall identify the data categorized by the analysis functionality.***

Table 7. (*continued*)

EXT_FDP_MON.1	
Family Behavior	Real-time monitoring of data leakage (EXT_ FDP_MON, monitoring) This family intends to monitor if the identified data is being leaked in real time.
Component description & structure	EXT_FDP_MON.1 Real-time monitoring of data leakage monitors the data and possible way of leakage in real time to prevent data leakage.
	EXT_FDP_MON.1 Real-time monitoring of data leakage Hierarchical to: No other components Dependencies: No dependencies ***EXT_FDP_ANL.1.1 The TSF shall ensure that the function to enforce the policy is invoked and performed successfully before the data is leaked.***
EXT_FDP.PRV.1	
Family Behavior	Prevention of data leakage (EXT_FDP_PRV, Prevention) This family intends to detect and prevent data leakage.
Component description & structure	EXT_FDP_PRV.1 Prevention of data leakage prevents data leakage from the protected system according to the policy set by an authorized administrator.
	EXT_FDP_PRV.1 Prevention of data leakage Hierarchical to: No other components Dependencies: EXT_FDP_MON.1 Real-time monitoring of data leakage ***EXT_FDP_PRV.1.1 The TSF shall perform the following in case potential security violation or data leakage is detected.*** ***a) Notify the authorized administrator***
Component description & structure	**b) Take actions as directed by the management of security functions** **c) Collect detailed information regarding the events of security violation** **d) [Assignment: Take actions in case of security violation in the protected system]** **EXT_FDP_PRV.1.2 The TSF shall store the following information about the actions taken.** **a) Actions taken and the result** b) **[Assignment: Information regarding the security violation in the protected system]**

3.5 Security Functional Requirements

The Security functional requirements substantiate the security objectives. Each security functional requirement must be related to one or more security objectives. These requirements are defined in CC part 2, and protection profile author just chooses and uses appropriate requirements. In addition, if the requirements defined in CC part 2 are not sufficient to demonstrate the security objectives, then, the protection profile author can refine and reinforce conditions in detail to established requirements. The security functional requirements for this paper are described in Table 8.

Table 8. The Security Functional Requireiments

Security Functional Class	***Components***	
Security audit	FAU_ARP.1	Security alarms
	FAU_GEN.1	Audit data generation
	FAU_GEN.2	User identity association
	FAU_SAA.1	Potential violation analysis
	FAU_SAR.1	Audit review
	FAU_SAR.3	Selectable audit review
	FAU_SEL.1	Selective audit
	FAU_STG.1	Protected audit trail storage
	FAU_STG.3	Action in case of possible audit data loss
	FAU_STG.4	Prevention of audit data loss
User Data Protection	FDP_ACC.1	Subset access control
	FDP_ACF.1	Security attribute based access control
	FDP_IFC.1	Subset information flow control
	FDP_IFF.1	Simple security attributes
	EXT_FDP_COL.1	Monitored data collection
	EXT_FDP_ANL.1	Monitored data analysis
	EXT_FDP_MON.1	Real-time monitoring of data leakage
	EXT_FDP_PRV.1	Prevention of data leakage
Identification and authentication	FIA_AFL.1	Authentication failure handling
	FIA_ATD.1	User attribute definition
	FIA_SOS.1	Verification of secrets
	FIA_UAU.2	User authentication before any action
	FIA_UAU.7	Protected authentication feedback
	FIA_UID.2	User identification before any action
Security Management	FMT_MOF.1	Management of security functions behavior
	FMT_MSA.1	Management of security attributes
	FMT_SMF.1	Specification of management functions
	FMT_SMR.1	Security roles
	EXT_FMT_STA.1	Data statistics of audit and leakage detection
Protection of the TSF	FPT_STM.1	Reliable time stamps
	FPT_TST.1	TSF testing
TOE Access	FTA_SSL.3	TSF-initiated termination

Our protection profile adopts EAL 4+ level in common criteria. Because DLP System is a critical information system and the result of attack can cause terrible confusion in society, we extend security assurance requirements to reinforce verification of implementation DLP system. Extended requirements are ADV_IMP.2, ATE_DPT.3, AVA_VAN.4.

4 Conclusion

Many companies are adopting a DLP system hoping that it can solve the desperate problem they are facing, risk of data leakage. This paper intends to suggest a baseline for introducing/evaluating a DLP system and eventually help protect data and prevent its theft or leakage. However, unlike the perimeter security, which is the essential aspect of IT security, data leakage protection is more about business than IT technologies because it deals with information assets, which is the most important thing to a company. That is, mere introduction of a DLP system does not prevent data leakage. For a complete prevention of data leakage, not only the IT team members but all company workers should aware their roles and responsibilities.

References

1. Lee, S., Shin, M.: Protection Profile for Software Development Site. In: Gervasi, O., Gavrilova, M.L., Kumar, V., Laganá, A., Lee, H.P., Mun, Y., Taniar, D., Tan, C.J.K. (eds.) ICCSA 2005, Part II. LNCS, vol. 3481, pp. 499–507. Springer, Heidelberg (2005)
2. Kanagasinghan, P.: Data Loss Prevention: SANS Institute InfoSec Reading Room (August 2008)
3. A Trend Micro White Paper: Addressing compliance Requirements for Privacy, Data Retention, and e-Discovery (March 2009)
4. The necessity of DLP Solutions: Osterman Research White Paper (June 2010)
5. SANS Institute: Understanding and Selecting a Data Loss Prevention Solution
6. IRONPORT SYSTMES with A Foreword By BRADLEY R. HUNTER: Data Loss Prevention Best Practices, Managing Sensitive Data in the Enterprise: A Message Media Publication
7. 21CSOFT; McAfee Total Protection for Data
8. McAfee; Application Control v5.0, Change Control v5.0, and Integrity Monitor v5.0 with McAfee Agent v4.5 and ePolicy Orchestrator v4.5 Security Target Document Version 0.6 (December 2010)
9. RSA: The Security Division of EMC RSA®Data Loss Prevention Suite v6.5 Security Target Document Version 0.7 (April 2009)
10. Aberdeen Group: The Cost-Based Business Case for DLP(June 2009)
11. RSA: RSA Solution Brief
12. Symantec, http://www.symantec.com
13. TREND MICRO, http://www.trendmicro.com
14. RSA, http://www.rsa.com
15. BlueCoat, http://www.bluecoat.com
16. Componentshop, http://www.componentshop.co.kr/isrc/mp/2009/08/dlp/dlp_01.html
17. WIKIPEDIA, http://en.wikipedia.org/wiki/DLP

Security Analysis on Digital Signature Function Implemented in PDF Software*

Sunwoo Park[1], Changbin Lee[1], Kwangwoo Lee[1], Jeeyeon Kim[1], Youngsook Lee[2], and Dongho Won[1,**]

[1] Information Security Group,
Sungkyunkwan University, Republic of Korea
{swpark,cblee,kwlee,dhwon}@security.re.kr, jeeyeonkim@paran.com
[2] Department of Cyber Investigation Police,
Howon University, Republic of Korea
ysooklee@howon.ac.kr

Abstract. Recently, electronic documents are deployed in many areas, thanks to their cost-efficiency and utility. However, trust concerns arise owing to the hardness of detecting document modification. To solve these concerns, many document processing software provide digital signature function. However, not much research was done to diagnose the security of implemented digital signature. Therefore, in this paper, we analyze the security of digital signature function implemented in PDF software including Adobe Acrobat, Nuance PDF Converter, and Foxit Phantom, and propose a list of recommendations for PDF software developers.

Keywords: Electronic document, digital signature, certificate, security analysis.

1 Introduction

Recently, green IT has become one of the primary issues globally. In conformance with this worldwide trend, using electronic document instead of paper document is encouraged for various purposes. However, though it reduces amount of wastepaper, electronic document has trust concerns. Modification of electronic documents cannot be detected easily. In addition, it is difficult to figure out whom the document was modified by. To solve these problems, digital signature is used in many document processors. By verifying digital signature, the integrity of electronic document can be checked, and the signer and time of sign can be acknowledged as well[1][2]. Therefore, digital signature capability is implemented in many document processors.

* This work was supported by grant RND project "A Study on visible digital signature of electronic document" of KISA.

This work was supported by Priority Research Centers Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2011-0018397).

** Corresponding Author.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 327–334, 2011.

PDF (Portable Document Format), developed by Adobe Systems in 1993, is a standard for document exchange, and is one of the most popular electronic document formats. There are many pieces of PDF software that provide digital signature function for PDF documents. However, not much research was done to diagnose the security of implemented digital signature. Throughout this paper, we analyze the digital signature function of existing PDF software to diagnose the security. Then we propose a list of recommendations that helps implementing digital signature function in a secure manner.

The rest of paper is organized as follows. In Section 2, we provide generic information on PDF to help understanding our paper. In Section 3, we give a detailed description of our analysis methods. Then, we show our analysis results in Section 4. In the following section, we propose a list of recommendations considering the vulnerabilities identified by our analysis. Finally, we conclude our paper in Section 6.

2 The PDF Standard

In January 2008, the ISO technical committee approved the final revised documentation for PDF 1.7 as the international standard ISO 32000-1[3]. '12.8 Digital Signatures' of ISO 32000-1 addresses signature dictionary. Signature dictionary contains information required for digital signature service (e.g. signature, contents, cert). Hence, PDF software should use signature dictionary to generate or verify digital signature. Especially, the name of handler which is special software module necessary for verifying digital signature is recorded on Filter field of signature dictionary. PDF software may substitute a different handler when verifying the signature, as long as it supports the specified SubFilter format. Example signature handlers are Adobe.PPKLite, Entrust.PPKEF, CICI.SignIt, and VeriSign.PPKVS. SubFilter describes the encoding of the signature value and key information in the signature dictionary. PDF software may use any handler that supports this format to validate the signature. PDF defined three SubFilter(adbe.x509.rsa_sha1, adbe.pkcs7.detached, and adbe.pkcs7.sha1). Our pieces of target software use adbe.pkcs7.sha1 SubFilter and adbe.pkcs7.detached SubFilter. If 'adbe.pkcs7.detached' is used for SubFilter, the original signed message digest over the document's byte range shall be incorporated as the normal PKCS#7 SignedData field. No data shall be encapsulated in the PKCS#7 SignedData field. Otherwise, if adbe.pkcs7.sha1 is used for SubFilter, the SHA1 digest of the document's byte range shall be encapsulated in the PKCS#7 SignedData field with ContentInfo of type Data. The digest of that SignedData shall be incorporated as the normal PKCS#7 digest[4].

3 Analysis Methods

To validate a digital signature, PDF software should verify the signature and the certificate which was used to create the signature. Thus, our analysis on digital signature function of PDF software is divided into two parts: analysis on digital signature verification, and analysis on certificate verification[5].

Table 1. Steps of Analysis Methods

Processes to be analyzed			Description
Digital Signature Verification			1. Verify the signature after digitally signing on an arbitrary electronic document. 2. Flip a bit of signature value. Then verify the signature again. If verification fails, we assume that the digital signature verification is being performed properly.
Certificate Verification	Certificate Chain		1. Check whether the application provides root certificate registration function. 2. Check whether the application shows certificate chain.
	Basic Information Verification	Signature Verification	1. Flip a bit of signature value field data of certificate. 2. Attempt to verify the signature. If verification fails, we assume that the signature verification is being performed properly.
		Status Verification	1. Generate a digital signature with a revoked certificate. If generation fails, we assume that the revocation check in digital signature generation process is being performed properly. 2. Attempt to verify a signature with a revoked certificate. If verification fails, we assume that the status verification is performed properly.
		Expiration Verification	1. Digitally sign with an expired certificate. If signing fails, we assume that the expiration check in digital signature generation process is being performed properly. 2. Attempt to verify a signature with an expired certificate. If verification fails, we assume that the expiration verification is performed properly. 3. Change the time of verifying PC to the time when certificate is valid. Then attempt to verify the signature. If verification fails, we assume that the time information used in expiration verification is reliable.

A certificate is verified to be valid only if it passes a series of certificate verification processes. Certificate verification process includes verification of certificate chain, and the verification of basic certificate information. The verification of basic certification information can be divided into signature verification, certificate status verification, expiration verification and policy verification[6][7]. However, in general, there are no limitations and constraints on certificate policy in the digital signature process. Therefore, we reviewed the software to check whether verification of certificate chain and basic information are implemented well. The analysis methods for each item are explained in Table 1.

4 Vulnerability Analysis

In this section, we analyze the security of digital signature function implemented on PDF software using methods described in the previous section. We analyzed three software, which are selected out of a list "Top Best PDF Software – Creator, Converter and Editor" published by CEOworld Magazine[8]. The pieces of target software are described in Table 2 below.

Table 2. List of Target Software

Company	Software	Version
Adobe	Acrobat X Pro	10
Nuance	PDF Converter	7.0
Foxit	Phantom	2.2.4

4.1 Adobe Acrobat

Digital Signature Verification. In Adobe Acrobat, we were able to see the signature verification result by clicking on the signature tag image. When we changed the signature value, Adobe Acrobat showed an error message that the document is modified and did not open the document. Namely, Adobe Acrobat verified the digital signature in a proper manner.

Verification of Certificate Chain. When we registered the root certificate as a trusted root certificate, the certificate status and the chain turned to be reliable and valid, respectively. Namely, Adobe Acrobat lets user to decide whether to trust the root certificate. Moreover properly verifies the certificate chain.

Verification of Basic Information of Certificate

Certificate Signature Verification. We tried to verify a signature after flipping a bit of certificate data field. The verification result showed that the signature is invalid, meaning that Adobe Acrobat verifies the certificate signature properly.

Certificate Status Verification. In Adobe Acrobat, we were able to generate a digital signature with revoked certificate. This means that Adobe Acrobat either does not verify certificate status, or performs the verification in an improper manner.

However, after performing several tests, we decided that upon verification of digital signature, Adobe Acrobat properly performs the certificate status verification. Adobe Acrobat brought OCSP (Online Certificate Status Protocol) information from AIA (Agency Information Access) field of the certificate to check the certificate status[9][10]. However, the verification failed several times due to incorrect data of AIA field and network problem.

Certificate Expiration Verification. Adobe Acrobat did not let us sign on the document when we attempted to sign with an expired certificate, by not allowing us to register the expired certificates. However, in the signature verification process, Adobe Acrobat did not provide any informative message that the certificate used in signing is expired; which makes users difficult to figure out. In addition, Adobe Acrobat used the time of verifying PC for certificate expiration verification in signature generation process and verification process. In this case, a malicious user can change the PC time, and successfully generate a signature or pass the signature verification. Hence, the certificate expiration verification process is vulnerable due to usage of unreliable time.

4.2 Nuance PDF Converter

Digital Signature Verification. In Nuance PDF Converter, we could see the verification result by clicking on the signature tag. When we changed the signature value, an error message popped up and the document was not loaded. That is, Nuance PDF Converter properly verified the digital signature.

Verification of Certificate Chain. Nuance PDF Converter did not provide a root certificate registration procedure. Moreover, it did not show relevant information. Hence, we could not check whether certificate chain verification is performed.

Verification of Basic Information of Certificate

Certificate Signature Verification. We tried to verify a signature after flipping a bit of certificate data field. The verification result showed that the signature is valid, meaning that Nuance PDF Converter does not verify the certificate properly.

Certificate Status Verification. Nuance PDF Converter generated a digital signature with revoke certificate. When we tried to verify that signature, Nuance PDF Converter showed a valid verification result. Namely, Nuance PDF Converter either does not verify the status of certificate, or verifies it improperly.

Certificate Expiration Verification. Nuance PDF Converter did not let us sign on the document when we attempted to sign with an expired certificate, by not allowing us to generate the digital signature. In addition, it showed an error message when we tried the verification with an expired certificate; which means that the certificate expiration verification is being performed properly. But, Nuance PDF Converter used the time of verifying PC for certificate expiration verification in signature generation process and verification process. In this case, a malicious user can change the PC time, and generate a signature or pass the signature verification successfully. Hence, the certificate expiration verification process of Nuance PDF Converter is vulnerable due to usage of unreliable time.

4.3 Foxit Phantom

Digital Signature Verification. In Foxit Phantom, we could see the verification result by clicking on the signature tag. When we changed the signature value, Foxit

Phantom showed a message that the document is modified and did not open the document. That is, Foxit Phantom properly verified the digital signature.

Verification of Certificate Chain. Foxit Phantom does not provide a root certificate registration procedure. Moreover, it did not show the certificate chain information. Hence, we could not check whether Foxit Phantom properly verifies certificate chain.

Verification of Basic Information of Certificate

Certificate Signature Verification. We tried to verify a signature after flipping a bit of certificate data field. The verification result showed that the signature is valid, meaning that Foxit Phantom does not verify the certificate signature properly.

Certificate Status Verification. Foxit Phantom generated a digital signature with revoke certificate. When we tried to verify that signature, Foxit Phantom showed a valid verification result. Namely, Foxit Phantom either does not verify the status of certificate, or verifies it improperly.

Certificate Expiration Verification. Foxit Phantom did not let us sign when we attempted to sign with an expired certificate. Instead, it showed a message warning that the certificate is an expired certificate and did not generate a signature value. However, even if digital signature is not generated, Foxit Phantom shows the image of digital signature on the document. This particular implementation flaw could possibly confuse users. In addition, Foxit Phantom used the time of verifying PC for certificate expiration verification in signature generation process and verification process. In this case, a malicious user can change the PC time, and generate a signature or pass the signature verification successfully. Hence, the certificate expiration verification process of Foxit Phantom is vulnerable due to usage of unreliable time.

4.4 Summary of Analysis Result

Table 3 provides summarized result of analysis on each software; Adobe Acrobat, Foxit Phantom, and Nuance PDF Converter. The result shows that all three software performs digital signature verification properly. However, Foxit Phantom and Nuance PDF Converter did not even verify the certificate basic information in a proper manner. This means that a malicious user can generate valid digital signature with a compromised certificate or with an invalid certificate. Therefore, the document is not trustworthy since the trust of electronic documents relies on the digital signature. In addition, Adobe Acrobat did verify the status of certificate for digital signature verification. However, Adobe Acrobat did not perform certificate status verification in digital signature generation process. This means that the software cannot prevent repudiation of malicious users[11].

Furthermore, upon verifying the expiration date of certificate, all three software use the time of verifying PC. Therefore, we can conclude that the certificate verification results are not reliable and the non-repudiation cannot be satisfied.

Table 3. Summarized result of analysis on each software

Software	Digital Signature Verification	Certificate Chain Verification	Signature Verification	Status Verification	Expiration Verification
Adobe Acrobat	O	O	O	▲	△
Foxit Phantom	O	?	X	X	△
Nuance PDF Converter	O	?	X	X	△

O : performs verification properly.
X : does not perform verification.
△ : performs verification insufficiently.
▲: performs verification only for digital signature verification process.
? : unknown.

5 Recommendation

Our analysis in the previous section has shown that PDF software either do not perform or insecurely perform the verification of certificate's basic information. For more reliable document verification system, developers should consider following to perform digital signature function of PDF software securely and usefully.

Digital Signature Verification. PDF software should provide a visible mark for users to figure out the presence of digital signature. The mark should be informative and easily seen by users, especially when the verification of digital signature fails.

Certificate Chain Verification. The trust of certificate chain mainly relies on the reliability of root certificate, and several root certificates are built in the operating system for convenience. However, there are numerous trustworthy root certificates which are not built in the system. Therefore, PDF software should have a function that helps users to register root certificates to help constructing a certificate chain. In addition, the certificate chain information should be easily seen by users, especially for informative purposes for failure of verification.

Certificate Signature Verification. PDF software should verify the signature of certificate used in digital signature to confirm the integrity. If the verification fails, PDF software should warn users of such verification failure, and inform the untrustworthiness.

Certificate Status Verification. PDF software should utilize either CRL (Certificate Revocation List) or OCSP (Online Certificate Status Protocol) to check the status of certificate for both digital signature generation and digital signature verification process. For the cases that real time certificate status is required, using OCSP is recommended[12][13][14]. If the certificate is revoked or invalid for any reason, the software should inform users that the digital signature is untrustworthy.

Certificate Expiration Verification. PDF software should provide a measure to prohibit generation of digital signature with an expired certificate for non-repudiation purposes. In addition, upon verification of digital signature, the expiration date should

be checked. If certificate is expired, the software should inform users that the digital signature is not reliable. PDF software should use reliable time value brought from a trustworthy server. Using timestamp token can be an effective measure.

6 Concluding Remarks

Recently, green IT has become one of the primary issues globally. In conformance with this worldwide trend, using electronic document instead of paper document is encouraged for various purposes. However, though it reduces amount of wastepaper, electronic document has trust concerns. Throughout this paper, we analyzed the digital signature function of existing software to diagnose the security. Then we listed several recommendations which may be helpful for implementing trustworthy digital signature function in document processing software.

References

1. Merkle, R.C.: A Certified Digital Signature. In: Brassard, G. (ed.) CRYPTO 1989. LNCS, vol. 435, pp. 218–238. Springer, Heidelberg (1990)
2. Bellare, M., Miner, S.K.: A Forward-Secure Digital Signature Scheme. In: Wiener, M. (ed.) CRYPTO 1999. LNCS, vol. 1666, pp. 431–448. Springer, Heidelberg (1999)
3. ISO 32000-1:2008. First Edition 2008-7-1: Document management - Portable document format - Part 1: PDF 1.7, http://www.iso.org/
4. Kaliski, B.: PKCS#7: Cryptographic Message Syntax (RFC 2315), RSA Laboratories (1998)
5. Merkle, R.C.: A Certified Digital Signature. In: Brassard, G. (ed.) CRYPTO 1989. LNCS, vol. 435, pp. 218–238. Springer, Heidelberg (1990)
6. Housley, R., Polk, W., Ford, W., Solo, D.: Internet x.509 public key infrastructure certificate and CRL profile. IETF RFC 3280 (2002)
7. Lee, Y., Ahn, J., Kim, S., Won, D.H.: A PKI System for Detecting the Exposure of a User's Secret Key. In: Atzeni, A.S., Lioy, A. (eds.) EuroPKI 2006. LNCS, vol. 4043, pp. 248–250. Springer, Heidelberg (2006)
8. CEOworld Magazine, http://ceoworld.biz/ceo/2010/04/13/
9. Lee, Y., Kim, I.J., Kim, S., Won, D.H.: A Method for Detecting the Exposure of OCSP Responder's Session Private Key in D-OCSP-KIS. In: Chadwick, D., Zhao, G. (eds.) EuroPKI 2005. LNCS, vol. 3545, pp. 215–226. Springer, Heidelberg (2005)
10. Myers, M., Ankney, R., Malpani, A., Galperin, S., Adams, C.: X.509 Internet PKI Online Certificate Status Protocol. IETF RFC 2560 (1999)
11. Gürgens, S., Rudolph, C.: Security Analysis of Efficient (Un-)fair Non-repudiation Protocols. Formal Aspects of Computing 17(3), 260–276 (2005)
12. Kocher, P.C.: On Certificate Revocation and Validation. In: Hirschfeld, R. (ed.) FC 1998. LNCS, vol. 1465, pp. 172–177. Springer, Heidelberg (1998)
13. Naor, M., Nissim, K.: Certificate revocation and certificate update. In: Proc. of USENIX Security Symposium, pp. 217–228 (1998)
14. Micali, S.: Certificate revocation system. United States Patent, US Patent 5,666,416 (1997)

Information Technology Security Evaluation Using CERT C Secure Coding Standard

Taeseung Lee[1], Kwangwoo Lee[1], Dongho Won[1], and Namje Park[2,*]

[1] Information Security Group,
School of Information and Communication Engineering,
Sungkyunkwan University, 300 Cheoncheon-dong, Jangan-gu,
Suwon, Gyeonggi-do 440-746, Korea
{tslee,kwlee,dhwon}@security.re.kr
[2] Department of Computer Education Teachers College,
Jeju National University, Korea
namjepark@jejunu.ac.kr

Abstract. IT products developed without due consideration of security issues have caused many security accidents over the last ten years. As a result, the importance of security in software development is increasing. It is important to ensure that no known vulnerabilities remain in the design, development, and test stage, in order to develop secure IT products. Even when an IT product is designed securely, various security vulnerabilities can occur, such as buffer overflow, if the general coding technique is used at the development stage. Therefore, the introduction of secure coding rules becomes most critical in developing a robust information security product. This paper proposes a method of applying a secure coding standard in the CC evaluation process. The proposed method is expected to contribute to improving the security of IT products in the CC evaluation process.

Keywords: Common criteria, Secure coding, CERT C Secure Coding Standard, Evaluation.

1 Introduction

The software defects, bugs and logic flaws are consistently the primary cause of commonly exploited software vulnerabilities. Through the analysis of thousands of reported vulnerabilities, security professionals have discovered that most vulnerability stem from a relatively small number of common software programming errors. By identifying the insecure coding practices that lead to these errors and educating developers on secure alternatives, organizations can take proactive steps to help significantly reduce or eliminate vulnerabilities in software before deployment [1].

In this manner, secure software programming is important to develop secure IT product. Recently, IT product is evaluated by the Common Criteria for Information

* Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 335–342, 2011.

Technology Security Evaluation (abbreviated as Common Criteria or CC)[3, 4, 5]. CC is an international standard (ISO/IEC 15408) for computer security certification. It provides computer system users can specify their security functional and assurance requirements, vendors can then implement and/or make claims about the security attributes of their products, and testing laboratories can evaluate the products to determine if they actually meet the claims. In other words, CC provides assurance that the process of specification, implementation and evaluation of a computer security product has been conducted in a rigorous and standard manner [2].

However, CC does not require secure coding in evaluation process. Therefore, the evaluators have difficulty to evaluate the secure implementation. To solve this problem, we have tried to adopt secure coding standard in the CC evaluation process.

2 CERT C Secure Coding Standard Overview

Secure coding rules are required to develop secure IT products. However, there is no international standard for secure coding, and many organizations draw up their own internal guidelines for secure coding. The CERT set up and distributed the CERT C Secure Coding Standard as an example of these coding rules. The CERT C Secure Coding Standard v1.0 was developed by the community, with the participation of about 320 technical experts, and was revised in May 2010 through continuous review.

The U.S. Department of Defense (DoD) demands compliance with secure coding rules as a part of the software assurance (SwA) system for commercial procurement products, and also demands compliance with the CERT C Secure Coding Standard. In addition, the chairman of the CERT submitted the CERT C Secure Coding Standard to the ISO/IEC JTC 1/SC 22/WG14 for the first time at the ISO London Conference in 2007, and discussion about the standard has continued ever since then. The CERT C Secure Coding Standard is processed by the ISO/IEC JTC1/SC22/WG14, and commercial analysis tools like Fortify, Klocwork, ROSE, and coverity have partially adopted the CERT C Secure Coding Standard.

The RFP of the U.S. DoD includes the “Application Security and Development Security Technical Implementation Guide” (hereafter referred to as “STIG”).

The DoD demands compliance with the coding standard for commercial products according to the above regulation, and the CERT developed the CERT C Secure Coding Standard for this purpose. In addition, representative commercial tools for source code analysis (fortify, coverity, so on.) have also adopted the CERT C Secure Coding Standard, besides the U.S. DoD. Therefore, the CERT C Secure Coding Standard can be applied to CC evaluation properly. For more detail information, please, refer to the [1].

3 Proposed Evaluation Methods for Secure Coding

3.1 Verifying Compliance with the ANSI C Standard Rules

The majority of commercial C language compiler tools check compliance with the standard during the compilation process. Therefore, the developer should check the C

language standard document supported by the commercial compilation tool identified by ALC_TAT.2. In addition, the developer should check whether any rule violates the standard rule or whether there is any additional rule besides the standard rule, by checking the user's guide for the commercial compilation tool. Then, the developer should check whether implementation expression of the TOE uses the rule in question.

If any exceptional or additional rule is included, besides the TOE implementation expression, it should be defined and the reason why the existing standard rules are not being conformed to should be presented. The developer's theoretical basis can be described in the evaluation submission document that corresponds to ALC_TAT.2, or ADV_INT.2. However, if the basis is described in one of either ALC_TAT.2 or ADV_INT.2, there should be a reference mark in the counterparty, as ALC_TAT.2 and ADV_INT.2 have a link. Figure 1 shows the procedure that verifies compliance with the ANSI C standard rule.

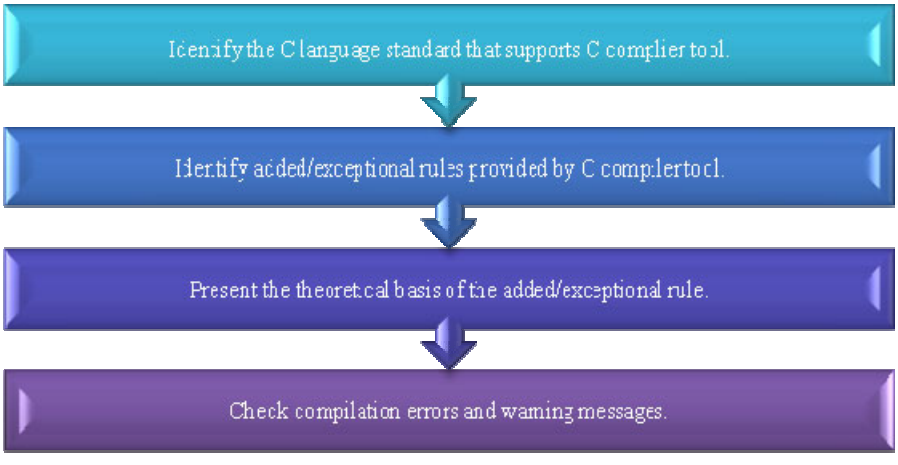


Fig. 1. Procedure for verifying compliance with the ANSI C standard

3.2 Verifying Compliance with the CERT C Secure Coding Standard

The TOE implementation expression may not correspond to all the rules or recommendations of the CERT C Secure Coding Standard. Therefore, the developer should comply with the CERT C Secure Coding Standard, and the evaluator should follow the following verification procedure. If any of steps 1 ~ 3 are taken, the evaluator can make the final table of the rules and regulations of the CERT C Secure Coding Standard for the TSF source code standard conformance target.

Figure 2 shows the procedure for verifying compliance with the CERT C Secure Coding Standard.

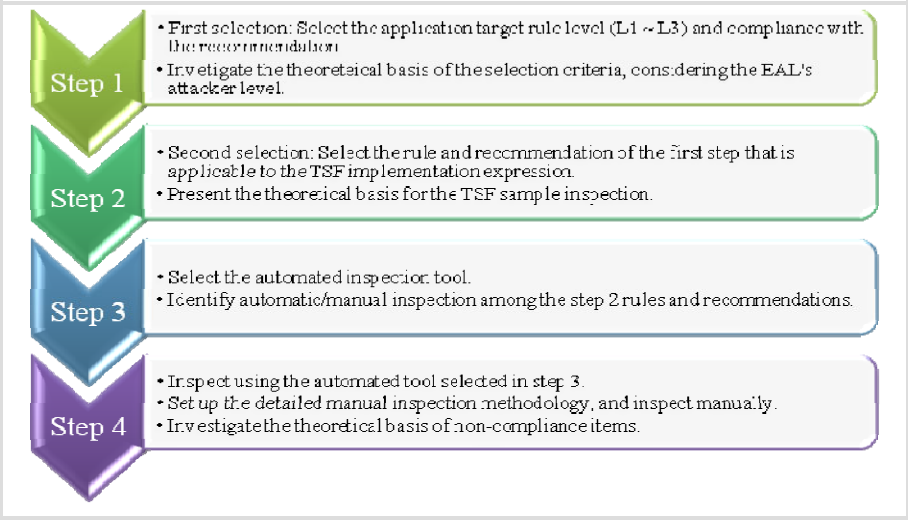


Fig. 2. Procedure for verifying compliance with the CERT C Secure Coding Standard

3.2.1 Step 1: First Selection of Rule/Recommendation Application Targets

The evaluator should choose the first selection criteria that fit into the product characteristics and evaluation environment, in order to filter out the first targets that will be applied to the TSF implementation expression among all the rules and recommendations of the CERT C Secure Coding Standard.

First Method. Selection by risk: Set the criteria grade of the risk determinants and calculate the risk, and then select the first rule/recommendation.

Second Method. Selection by risk determinant grade: Select the first rule/recommendation by several combinations of three determinants.

Combination 1) Rules/Recommendations that have medium or over severity ∩ probable or higher likelihood

Combination 2) Rules/Recommendations that have medium or over severity ∪ probable or higher likelihood

Combination 3) Rules/Recommendations that require low improvement cost ∩ and high severity

In this paragraph, the risk is selected as the first selection criteria for the rule/recommendation that will be applied to the TSF implementation expression, considering the case of general software.

The risk of the rule can be set, considering the attacker level and product type of AVA_VAN.# that is included in the EAL. The risk of the recommendation can be set to be the same as the risk of the rule, or it can be set differently.

3.2.2 Step 2: Second Selection of Rule/Recommendation Application Targets

The evaluator should establish the proper second selection method, depending on the TSF characteristics. For example, the evaluator should arrange the source code by requesting the developer to remove any duplicated functions, the commented and inactive code, and any function that is not invoked at all during TSF execution, using the ADV_INT.2-3 work unit. The rules and recommendations that can be applied to the TSF should be selected from among the first selection rules and recommendations at the first step, by analyzing the source code. The example of the evaluator's second selection method can be summarized as follows.

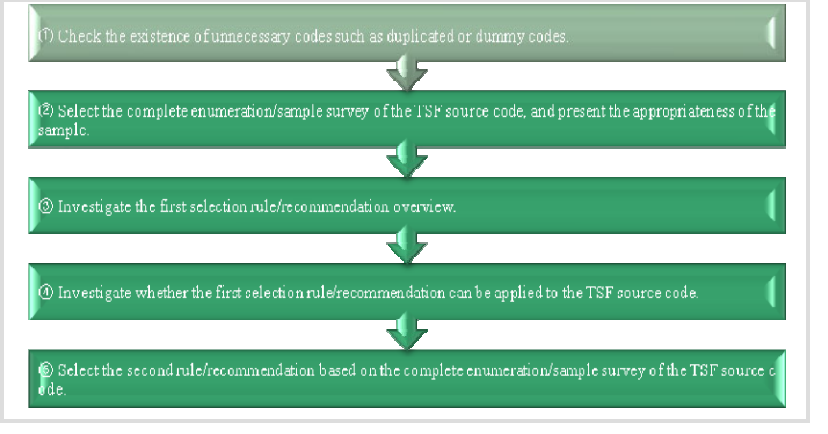


Fig. 3. Second selection method

The evaluator should present the reasonableness of the sample selection, if the sample survey is conducted on the TSF source code. Then, the evaluator should check whether the TSF source file contains the word corresponding to the reserved words of the C language (command, function, data structure, etc.) specified in the non-compliant code example of the rule and recommendation, in order to determine whether the first selection rule and recommendation can be applied to the TSF source code. A compilation tool or an automated tool like Source Insight can be used for searching.

3.2.3 Step 3: Identification of the Automatic/Manual Inspection Method

The automated tool can be selected as a tool that checks whether the TSF source code complies with the rule and recommendation selected in step 2. The evaluator can perform manual inspection to determine compliance with all the rules and recommendations, but it is inefficient and inaccurate. The evaluator may provide the second selection result, so that the developer can select an automated inspection tool that fits into the characteristics of the TSF source code. Figure 5 shows the number of rules/recommendations detected by the individual automated tool out of 243 possible rules and recommendations.

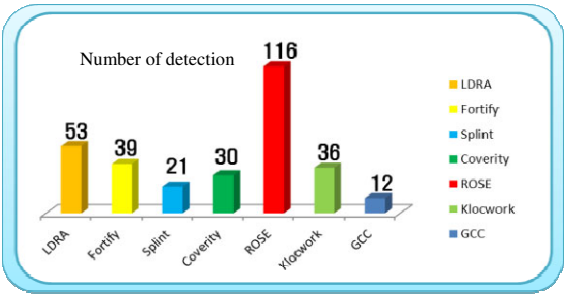


Fig. 4. The comparison of automated detection execution

In the above figure, ROSE cannot be the efficient tool in any case. For example, ROSE can detect the largest number of rules/recommendations. However, Fortify can detect more rules/recommendations when applied to the TOE. Therefore, the most efficient detection tool can vary, depending on the second selection result of the TSF source code application target in the second step.

3.2.4 Step 4: Checking Compliance with Rules/Recommendations

The automated tool is used for inspection in the third step. The allowable level should be selected for the tool, and the evaluator should request the developer to modify the rule, if it deviates from the allowable level. If the rule cannot be modified, the evaluator should request the developer to present the theoretical basis for the rule/recommendation that deviates from the allowable level.

Rule	Summary		Detailed methodology of manual inspection
PRE 01-C	Parentheses should be used for the argument of extended strings, because macro functions can cause side-effects due to the additional operation of extended strings, depending on the passed argument. However, arguments are separated by a comma, but parentheses may not be used.		Identify the #define statement that includes arithmetic operation expression among extended strings, by searching all statements defined with the #define statement, and then check whether any #define statement is missing parentheses.

Fig. 5. Establishment of detailed manual inspection methodology

An individual detailed inspection methodology should be established for each rule/recommendation for the manual inspection items in the final list of the TSF application target, which was created during the third step. For example, a methodology that stipulates how the evaluator should conduct an inspection should be established.

The evaluator should investigate whether the TSF source code complies with RPE01-C according to the detailed manual inspection methodology. If there is a

#define statement without parentheses, the evaluator should request the developer to modify it. If the developer cannot modify the statement due to rational reasons, the evaluator should demand the theoretical basis and check its validity.

Based on the automatic and manual inspection described above, the following conclusion can be drawn.

Conclusion. The TSF complies with the CERT C Secure Coding Standard (May 2010) by conforming to the rules of the risk L# or above, and the recommendations of the risk level L# or above. Therefore, the TSF has no vulnerability due to source code defects at the AVA_VAN.# level.

Fig. 6. The evaluation results of automated and manual inspection

4 Simulation Results

The procedure for verifying compliance with the CERT C Secure Coding Standard was applied to the mock TOE (PC firewall application program) of the information security product in this study. The verification procedure is as follows (the details of which are not covered by this paper due to the limitation on its length):

Step 1) The risk was selected, based on the TOE overview information. The first rule/recommendation was selected that will be applied to the TSF from among the entire rule, based on the selected risk (L1 ~ L3).

Step 2) The second rule/recommendation corresponding to the TSF implementation expression was selected from the first selection group.

Step 3) After identifying the automated inspection tool selected by the applicant, automated and manual inspection of the individual rule and recommendation was determined.

Step 4) The second group was inspected using the automated inspection tool. The evaluator set up the detailed inspection methodology for the rule/recommendation that cannot be inspected by the automated tool, and conducted an inspection.

Conclusion. The TSF complies with the CERT C Secure Coding Standard (May 2010) by conforming to the rules of the risk 1# or above, and the recommendations of the risk level L1 or above. Therefore, the TSF has no vulnerability due to source code defects at the AVA_VAN.4 level.

Fig. 7. The evaluation simulation results of proposed methodology

5 Conclusion

CC does not require secure coding in evaluation process. Therefore, the evaluators have difficulty in assuring the security in software or firmware. To solve this problem, we adopted the CERT C secure coding standard in the ADV_INT2-4 of CC evaluation process. The proposed methodology can be used to increase reliability of the CC evaluation and security of IT product.

References

1. Lee, S.-y., Shin, M.-c.: Protection Profile for Software Development Site. 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. 499–507. Springer, Heidelberg (2005)
2. Common Criteria, Common Criteria for Information Technology Security Evaluation;part 1: Introduction and general model, Version 3.1 R1, CCMB-2006-09-001 (September 2006)
3. Common Criteria, Common Criteria for Information Technology Security Evaluation;part 2: Security functional components, Version 3.1 R2, CCMB-2007-09-002 (September 2007)
4. Common Criteria, Common Criteria for Information Technology Security Evaluation;part 3: Security assurance components, Version 3.1 R2, CCMB-2007-09-003 (September 2007)
5. Lee, K., Won, D., Kim, S.: A Secure and Efficient E-Will System Based on PKI. Information - An International Interdisciplinary Journal, International Information Institute 14(7), 2187–2206 (2011)
6. Lee, Y., Kim, S., Won, D.: Enhancement of two-factor authenticated key exchange protocols in public wireless LANs. Elsevier Computers and Electrical Engineering 36(1), 213–223 (2010)
7. Lee, K., Lee, Y., Won, D., Kim, S.: Protection Profile for Secure E-Voting Systems. In: Kwak, J., Deng, R.H., Won, Y., Wang, G. (eds.) ISPEC 2010. LNCS, vol. 6047, pp. 386–397. Springer, Heidelberg (2010)
8. Park, N., Song, Y., Won, D., Kim, H.: Multilateral Approaches to the Mobile RFID Security Problem Using Web Service. In: Zhang, Y., Yu, G., Hwang, J., Xu, G. (eds.) APWeb 2008. LNCS, vol. 4976, pp. 331–341. Springer, Heidelberg (2008)
9. Park, N., Kwak, J., Kim, S., Won, D., Kim, H.: WIPI Mobile Platform with Secure Service for Mobile RFID Network Environment. In: Shen, H.T., Li, J., Li, M., Ni, J., Wang, W. (eds.) APWeb Workshops 2006. LNCS, vol. 3842, pp. 741–748. Springer, Heidelberg (2006)
10. Park, N., Kim, H., Kim, S., Won, D.: Open Location-based Service using Secure Middleware Infrastructure in Web Services. 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. 1146–1155. Springer, Heidelberg (2005)
11. Park, N.: Security Scheme for Managing a Large Quantity of Individual Information in RFID Environment. In: Zhu, R., Zhang, Y., Liu, B., Liu, C. (eds.) ICICA 2010. CCIS, vol. 106, pp. 72–79. Springer, Heidelberg (2010)
12. Park, N., Song, Y.: Secure RFID Application Data Management Using All-Or-Nothing Transform Encryption. In: Pandurangan, G., Anil Kumar, V.S., Ming, G., Liu, Y., Li, Y. (eds.) WASA 2010. LNCS, vol. 6221, pp. 245–252. Springer, Heidelberg (2010)
13. Park, N.: Secure UHF/HF Dual-Band RFID: Strategic Framework Approaches and Application Solutions. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 488–496. Springer, Heidelberg (2011)
14. Park, N.: Secure Data Access Control Scheme Using Type-Based Re-encryption in Cloud Environment. In: Katarzyniak, R., Chiu, T.-F., Hong, C.-F., Nguyen, N.T. (eds.) Semantic Methods. SCI, vol. 381, pp. 319–327. Springer, Heidelberg (2011)
15. Park, N., Song, Y.: AONT Encryption Based Application Data Management in Mobile RFID Environment. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010. LNCS (LNAI), vol. 6422, pp. 142–152. Springer, Heidelberg (2010)

USN Middleware Access Control of Sensor Network and Selective Encryption of Information

Taeseung Lee[1], Dongho Won[1], and Namje Park[2,*]

[1] Information Security Group,
School of Information and Communication Engineering,
Sungkyunkwan University, 300 Cheoncheon-dong, Jangan-gu,
Suwon, Gyeonggi-do 440-746, Korea
{tslee,dhwon}@security.re.kr
[2] Department of Computer Education Teachers College,
Jeju National University, Korea
namjepark@jejunu.ac.kr

Abstract. This paper is to prepare guidelines for providing security functions that protect USN middleware system resources from malicious nodes by providing authentication, authorization and confidentiality for middleware connection nodes (applications/manager, sensor network) to make USN middleware safe.

Keywords: Sensor Network, Access Control, Encryption, USN.

1 Introduction

The USN middleware classifies two groups. One is USN application or administrator (consumer). Simply, The document calls it as an application. Application uses USN middleware to control USN infrastructure and acquire sensing information (raw sensing information or processed sensing information). The other group of users is USN infrastructure such as wireless/wired sensor network, RFID, mobile RFID, and IP-USN, etc. Simply, call them as a sensor network.

Applications use USN middleware to control sensor networks and collect sensing information from the sensor networks connected to the USN middleware. Applications send queries to USN middleware to acquire raw sensing value and/or processed information. Information which are integrated and derived from raw sensing data from a(or multiple) sensor network(s). The USN middleware interprets application requests and sends requests to various sensor networks in each sensor network comprehensible ways. Multiple applications can share the sensing information through USN middleware. USN middleware can provide raw sensing value from sensor networks. In addition, USN middleware integrates several raw sensing values from different sensor networks and even more provide processed information from several

* Corresponding author(namjepark@jejunu.ac.kr).

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 343–352, 2011.

sensing values and legacy data. Furthermore, USN middleware can derive processed information from raw sensing data, historical data and legacy data using mining technology, context-aware technology, and event processing technology. Application can control sensor networks which are connected to USN middleware. Application may activate/deactivate some kinds of actuators, change sensor network topology, or even change application running on sensor node dynamically. Usually, sensor network is powered by battery. And the devices such as sensor node, sink node, gateway are not cheap yet. Therefore, applications have to manage sensor network in a cost-effective way.

Sensor networks use USN middleware to provide sensing values to the applications. Sensor network provides its sensing value as response to the request or without explicit request. Usually, sensor network is used for environmental surveillance. For example, Sensor Web[7] led by NASA JPL(Jet Propulsion Laboratory), has been used to implement a global surveillance program to study volcanos. Sensor network senses environmental parameters such as temperature, humidity, pressure, etc. The way of sensing is usually periodic with some specific interval and lifetime. Often it responds just one time on receiving the request from application. USN middleware classify the queries into 4 groups. They are Instant Query, Continuous Query, Instant Query with Condition and Continuous Query with Condition. “Instant Query” means sensor network responds only one time at receiving the request from application. “Continuous Query” means the query such as “get temperature every 30 minutes during 30 days.” “Instant Query with Condition” means a kind of Instant Query restricting the response. For example, “get temperature if sensed temperature is over 30oC.” “Continuous Query with Condition” means a kind of Continuous Query restricting the response. For example, “get temperature every 30 minutes during 30 days if sensed temperature is over 30oC.”

From a USN middleware viewpoint, sensor networks are information providers. Sensing information flowing into USN middleware is flowing into several applications. Therefore, the genuineness of sensing information is very crucial to USN middleware and USN application.

2 Overview of USN Middleware System Security

The USN middleware platform provides sensor data mining and autonomic monitoring functions besides the connection function for a plurality of foreign sensor networks and the function of providing data for sensing data processing and application from these. Of these, the USN middleware security component that is to be dealt with in this paper has an object to protect USN middleware against external attack and sensing data provided through the USN middleware, and such information protection has an object to provide encryption of sensing information that requires access control and confidentiality for USN middleware.

USN middleware is an "information provider" that acquires raw data from USN infrastructure and provides data to the node that needs it after going through refining and processing works of a given pattern. If USN middleware is attacked, wrong data could not only make extensive application services malfunction but also throw into

confusion the USN infrastructure managed through the USN middleware. Therefore, security has a very great importance in USN middleware.

Protection of USN middleware as a 'sensor network' administrator as well as an 'information provider' is very important. If USN middleware is contaminated, wrong data could flow into the application along with proper data to contaminate the whole application system, and the sensor network control function by the USN middleware could be contaminated by a malicious application service or administrator so that all the managed USN resources would be plunged in confusion.

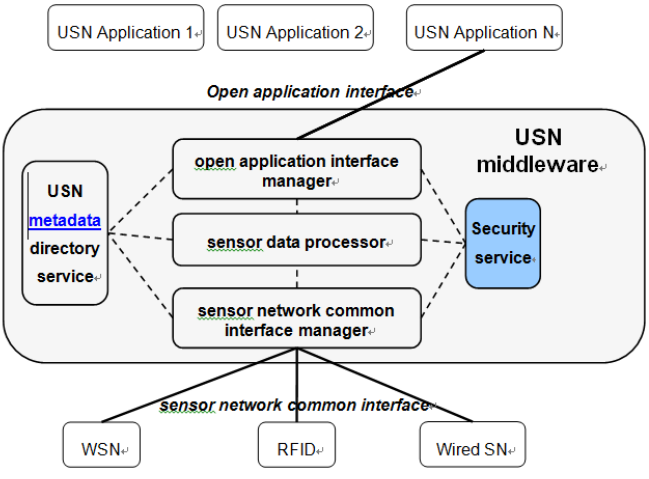


Fig. 1. Functional model of USN middleware

2.1 Security Threats by Applications

Application uses USN middleware to access sensor networks. As mentioned above, sensor networks are usually powered by batteries. It is the critical obstacle in USN industry. And the devices such as sensor nodes, sink nodes, gateways are not cheap until now. Therefore cost-effect usage of sensor network is very importance factor. Followings are possible security threats over USN middleware caused by applications.

- Authorized Sensing Information Acquisition: Unauthorized user can acquire various sensing data from multiple sensor networks through USN middleware. In this case, sensor networks may work more for unauthorized users than authorized users.
- Sensor Network Disruption: Application can control sensor network through USN middleware. By using USN middleware, application can invoke some kinds of actuators and change sensor networks configuration. By doing so, sensor networks may be configured in a wrong way. Sometimes, actuator such as alarm may cause some problems.
- Derive DoS attack: If unauthorized user sends sensing information request to the specific sensor network continuously, then that sensor network is going to be losing

all batteries. Ultimately, this sensor network is supposed to deny all requests from authorized users. This kind of DoS attack is very easily happened without any authentication and authorization mechanism support. The computing capability and storage available are very different among sensor networks. Some sensor network can' run multiple operations at the same time. In this case, request arrived late may be ignored or have to be queued. If unauthorized application requests sensing data and authorized application requests the same type of sensing data from the same sensor network later, then, authorized request may be ignored or queued until unauthorized request processing is over. Queued request may be discarded. This kind of DoS attack is also crucial to sensor network. That is because, sensor networks usually are powered by batteries, so they have lifetime.

2.2 Security Threats by Sensor Networks

Sensor networks use USN middleware to provide sensing value. If unidentified malicious sensor networks flow sensing information into USN middleware, then all applications which use that information are going to be corrupted. Followings are such possible security threats to be happened by malicious sensor networks

- Wrong Sensing Value Injection: In the USN middleware architecture, every application gets information from USN middleware. Those kinds of information are either raw sensing value or derived/integrated information from those raw ones. One of characteristics of using USN middleware is information-sharing among several applications through USN middleware, conditioned that they want sensing values from the same sensor networks. Fig. 3 depicts the wrong sensing value injection scenario. In this scenario, both of App #1 and App#2 use temperature values of SN#1. If SN#1 flow wrong values into the USN middleware, then App#1 and App#2 are disrupted together. It may cause activation of fire-alarm wrongly, or activate air conditioner at very low temperature.

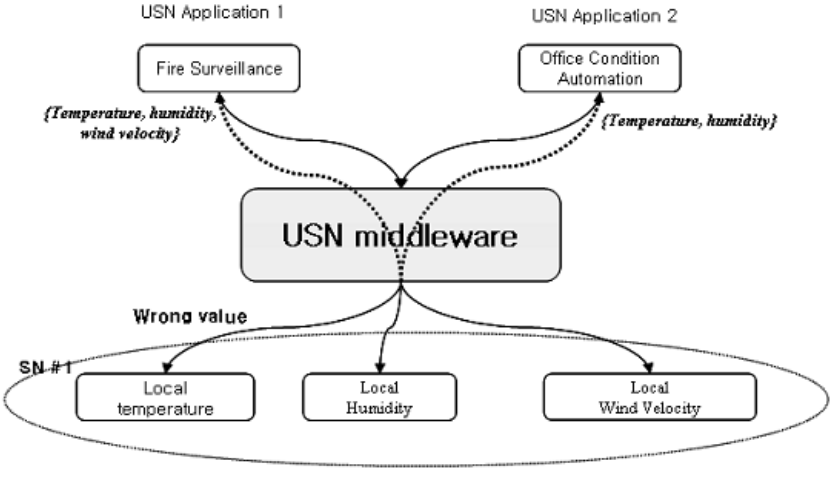


Fig. 2. Wrong Sensing Value Injection Scenario

2.3 Threat Caused by External Object

There are two open paths may be intruded by an attacker. One is between application and USN middleware and the other is between USN middleware and sensor network. Especially, the communication path between USN middleware and sensor network is wireless. That is easy to be attacked. Followings are possible attacks to be happened by exploiting two open paths.

- Eavesdropping: If attacker places between application and USN middleware, then it can obtain sensing information freely, without any charge. That is applied to the path between USN middleware and sensor network, too.

- Replay Attack: If attacker locates between application and USN middleware, he can obtain commands from application. After he acquires some specific command from an application, he resends the same command later again. Then USN middleware resends the same command to the same sensor network. In this way, attacker may exhaust the power of the sensor network and it cause DoS from that sensor network. Or, attacker obtains sensing value from USN middleware. Now, he intercepts command to be sent later, and then he reuses the sensing value obtained from the previous command and doesn' send the command to USN middleware. In this case, an application can' detect real situation. Therefore it can' deal with emergent events appropriately. Above cases are applied to the interface between USN middleware and sensor network, too. In this case, the communication path is wireless. Therefore, attacker can obtain message more easily than the above cases.

- Man-in-the-Middle Attack: Attacker may locate itself on the path between application and USN middleware and between USN middleware and sensor network. If attacker intercepts correct message from any one (app, USN MW, SN), he would change the contents of the message and send to the designated one. In this way, any object in USN middleware architecture can be failed easily.

3 Secure USN Middleware Reference Model

The function provided by USSC is to forward class to SNCIC and OAPIC so as to be used. In classifying the users, USSC defines through OAPIC the application (App) and administrator (Admin) that use USN middleware, and defines through SNCIC various adapters of USN infrastructure that access USN middleware as user. Especially in the USN infrastructure that accesses USN middleware through SNCIC, that is, all of the wireless/wired sensor network, RFID reader, and IP-USN nodes, are connected through the adaptors that provide protocol change and additional functions.

The internal blocks composing the USSC are as follows.

- Profile managing module (application, administrator, adaptor): In the inside are stored profiles for application, administrator and adaptor, and in the profile is stored information to be used for authentication, authorization and encryption of these.

- Profile setting module: Module for storing and applying administrator's settings in linkage with management tool (UI).
- USSC interface: Interface module of USSC, which OAPIC and SNCIC use to have security service provided, carries out the functions of authentication, authorization and encryption.
- Security processing module: Composed of authentication/authorization/encryption processing modules, it carries out authentication/authorization/encryption functions by using the processing module in each profile using the information set in the profile.
- Authentication session managing module: It carries out the role of managing authentication sessions and is in charge of managing the authentication tokens for the authenticated objects and the keys used in password.
- Authentication processing module: Module for processing authentication using user token file
- Authorization processing module (resource management and access control setting): Module for authorization processing based on user profile
- Password processing module: Module for processing passwords using user profiles
- Security library module: Library comprising security algorithm used in encoding and decoding, PKC (Public Key Certificate) processing module, and security protocol (TLS, etc.)
- Key management module: Module for managing the generation/distribution of the keys to be used in encryption
- Authentication token management module: Module for managing authentication tokens for the objects that succeeded in authentication
- Database Manager: In charge of connection and interface with DB that stores USSC policy
- User profile and security policy storage: Place for storing user profiles and security policy, which uses DB that is used in USN middleware.

3.1 Security Access Method of USN Middleware System

There are two methods of distributing parameters and algorithms necessary for authentication, authorization and encryption.

- Method whereby end nodes share each other offline
  * Since necessary nodes have necessary information in advance, it has an advantage that operation for sharing is not necessary.
  * It has a disadvantage that scalability is decreased in terms of maintenance.
  * It has another disadvantage that security is completely broken once information has leaked out since the same value is used continuously although there are no threats of leakage of information necessary for authentication/encryption.

- Method of distributing information online
  * Excellent in scalability since it is generated as necessary without sharing in advance.

* Distribution algorithm is necessary separately, and operation cost is needed.
* Takes a lot of operation cost as it is generated as necessary, while rather safe in terms of information protection.

3.2 Security Policy Scheme

The realization portion of security policy is IFA-M and IFM-S of Fig. 2. Namely, they are the portion of application service and USN middleware interface (IFA-M) and the portion of USN middleware and sensor network interface (IFM-S). Policy formulation for authentication, authority check and encryption of these and necessary profile management are to be processed by USSC. Namely, OAPIC or SNCIC becomes an intermediate node at the time of processing authentication and authority check, and actual processing is to be done by USSC. In the case of encryption, OAPIC/SNCIC are to do actual processing.

- In consideration of performance, it can be considered that each component can process by caching for cachable information.
- Whether to develop by using exclusive security server like AAA server, or USSC will do all the functions of policy and key distribution, will be specified by considering at the design stage.

USSC is to maintain/manage profile for application service and profile information on legal sensor network. At this time, the information is to use DB that exists in USN middleware. USN middleware policy for authentication, authority check and encryption is to be maintained by documentation. It should be made possible that such content is confirmed through USN middleware integrated GUI, and that whether each function is used or not is selectively activated. The security function is designed in such a way that it can be made active or inactive through integrated UI.

3.3 Security Component Developing Method

1) Access Control and Encryption for Application Service and Administrator(IFA-M)
For USN middleware protection for application service and administrator, use the OAPIC that is provided at present, expand the necessary portion, and utilize the authentication method that was carried out independently by OAPIC. Correction is possible as necessary.

Access control for application service and administrator and the method of selective encryption of information are as follows. The accessers of application level to USN middleware are application service and administrator. Control of access control to USN middleware by malicious users and encryption of information should be provided. The information protection function is provided by the following items.

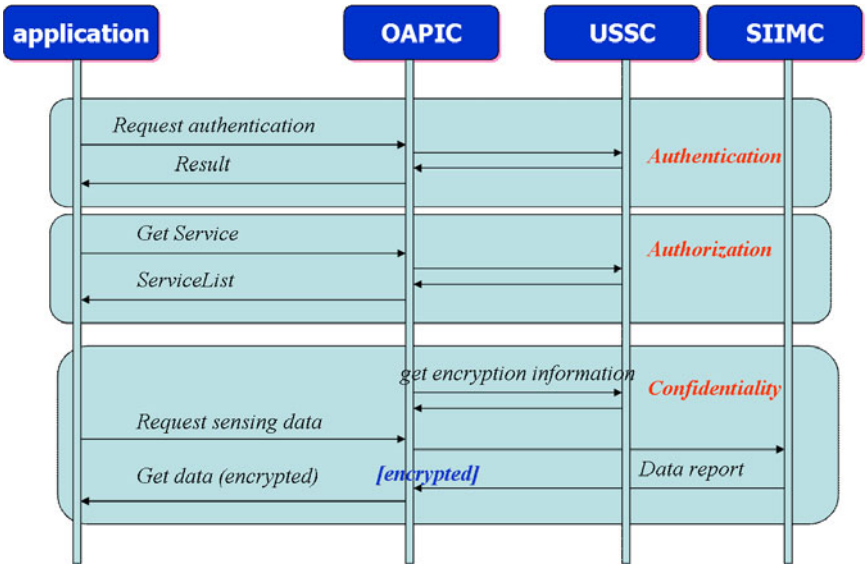


Fig. 3. Flow chart of process between application and USN middleware with Security

The detailed logging procedure is as follows.

- Control of access to USSC server.
- Accessible service list authority control [udsc, siimc, etc].
- Selective sensing data encryption
- User profile management
- Authorization for notification URL: If the application wants a separate notification URL, check whether it is possible URL.

2) Middleware Access Control of Sensor Network and Selective Encryption of Information (IFM-S)

USN middleware collects management and information on the sensor network to which information is to be provided, and provides them to the application. The function of control of access to USN middleware should be provided to prevent malicious sensor network from connecting to the USN middleware to provide infected bad information. An encryption function for information that is selectively flowed in should be also provided. And another function should be provided for connecting to USN middleware dynamically through authentication of sensor network that is not already registered in UDSC.

- Authentication possible when connected
  . SN that is already registered in UDSC
  . SN that is not already registered in UDSC: Authentication possible when connected to dynamic sensor network
- Function of selective encryption of sensing data

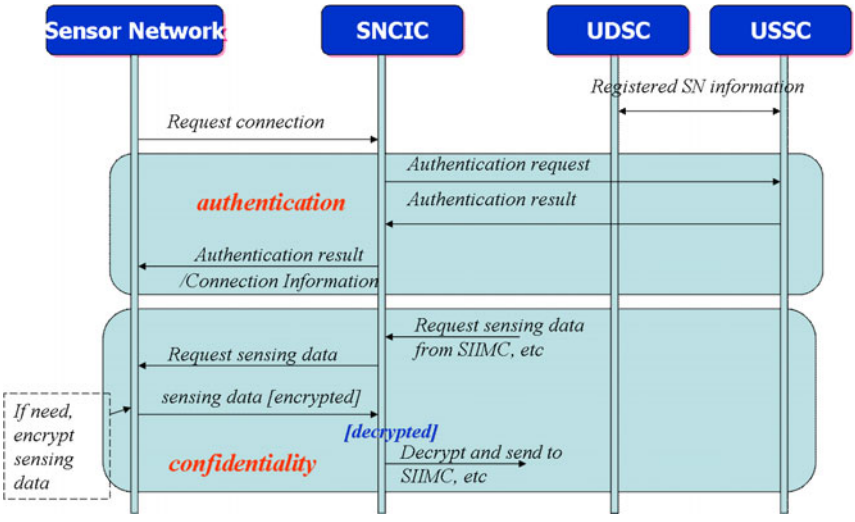


Fig. 4. Flow chart of process between USN middleware and sensor network with Security

4 Conclusions

USN middleware is the system to provide useful environmental information around the world, such as local temperature, local humidity, and local wind velocity, etc. By using that information, many kinds of automated environmental service provisioning can be realized such as automatic fire surveillance system, automatic ocean environmental supervision system, and u-silvercare system, etc. But until now, the maintenance cost of various sensor networks is not that cheap. Therefore, sensor network needs to be operated effectively to be lasting for a long time. To run various kinds of USN applications and sensor networks effectively, USN middleware is the one indispensable. USN middleware provides developmental and operational environment to develop USN applications easily and to manage sensor networks securely and cost effectively. But without consideration on security mechanism, whole USN computing society may be failed easily.

This paper is to prepare guidelines for providing security functions that protect USN middleware system resources from malicious nodes by providing authentication, authorization and confidentiality for middleware connection nodes (applications/manager, sensor network) to make USN middleware safe.

Acknowledgments. This paper is extended from a presentation paper presented at Geneva Meeting (“Proposal for a new work item on USN middleware security”, Q.9/17, WP2), GENEVA (2008). The author is deeply grateful to the anonymous reviewers for their valuable suggestions and comments on the first version of this paper.

References

1. Li, S., Lin, Y., Son, S.H., Stankovic, J.A., Wei, Y.: Event Detection Services Using Data Service Middleware in Distributed Sensor Networks. Thesis, University of Virginia (2003)
2. Kim, M., Lee, Y.: Design of Middleware for Pervasive Wireless Infrastructure. In: International Ubiquitous Conference (2006)
3. Yao, Y., Gehrke, J.: The Cougar Approach to In-Network Query Processing in Sensor Networks. Thesis, Cornell University (2002)
4. Shen, C.-C., Srisathapornphat, C., Jaikaeo, C.: Sensor Information Net-working Architecture and Applications. IEEE Personel Communication Magazine, University of Delaware (2001)
5. Yoon, M.Y., Kim, M.J., Park, N.J., Choi, D.H., Jung, H.: Proposal for a new work item on USN middleware security. Q.9/17, WP2 (2008)
6. Park, N., Song, Y., Won, D., Kim, H.: Multilateral Approaches to the Mobile RFID Security Problem Using Web Service. In: Zhang, Y., Yu, G., Bertino, E., Xu, G. (eds.) APWeb 2008. LNCS, vol. 4976, pp. 331–341. Springer, Heidelberg (2008)
7. Park, N., Kwak, J., Kim, S., Won, D., Kim, H.: WIPI Mobile Platform with Secure Service for Mobile RFID Network Environment. In: Shen, H.T., Li, J., Li, M., Ni, J., Wang, W. (eds.) APWeb Workshops 2006. LNCS, vol. 3842, pp. 741–748. Springer, Heidelberg (2006)
8. Park, N., Kim, H., Kim, S., Won, D.: Open Location-based Service using Secure Middleware Infrastructure in Web Services. 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. 1146–1155. Springer, Heidelberg (2005)
9. Park, N., Kim, S., Won, D.: Privacy Preserving Enhanced Service Mechanism in Mobile RFID Network. ASC, vol. 43, pp. 151–156. Springer, Heidelberg (2007)
10. Park, N.: Security Scheme for Managing a Large Quantity of Individual Information in RFID Environment. In: Zhu, R., Zhang, Y., Liu, B., Liu, C. (eds.) ICICA 2010. CCIS, vol. 106, pp. 72–79. Springer, Heidelberg (2010)
11. Park, N., Song, Y.: Secure RFID Application Data Management Using All-Or-Nothing Transform Encryption. In: Pandurangan, G., Anil Kumar, V.S., Ming, G., Liu, Y., Li, Y. (eds.) WASA 2010. LNCS, vol. 6221, pp. 245–252. Springer, Heidelberg (2010)
12. Park, N.: Secure UHF/HF Dual-Band RFID: Strategic Framework Approaches and Application Solutions. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 488–496. Springer, Heidelberg (2011)
13. Park, N.: Secure Data Access Control Scheme Using Type-Based Re-encryption in Cloud Environment. In: Katarzyniak, R., Chiu, T.-F., Hong, C.-F., Nguyen, N.T. (eds.) Semantic Methods. SCI, vol. 381, pp. 319–327. Springer, Heidelberg (2011)
14. Park, N., Song, Y.: AONT Encryption Based Application Data Management in Mobile RFID Environment. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010. LNCS (LNAI), vol. 6422, pp. 142–152. Springer, Heidelberg (2010)

Enhanced Code-Signing Scheme for Smartphone Applications*

Inkyung Jeun, Kwangwoo Lee, and Dongho Won**

Sungkyunkwan University,
300 Cheoncheon-dong, Jangan-gu, Suwon, Gyeonggi-do, 440-746 Korea
{ikjeun,kwlee,dhwon}@security.re.kr

Abstract. Recently, the number of incidents by malicious codes designed to suspend services and abuse personal information has grown rapidly, and the installation of applications on smart phones has emerged as one of the most common ways by which such malicious codes are spread. Anti-virus programs can be used to curb the spread of such codes, but these have limitations in terms of speed and efficiency. Accordingly, we need to strengthen the safety of application distribution and verification procedures in order to prevent the spread of malicious codes. To this end, this paper examines the problems of existing application distribution procedures, and suggests an enhanced code-signing scheme using the public key infrastructure (PKI) certificate for an application distribution method. It offers improved reliability and security by using code signing technology to secure the integrity of software and developer authentication functions.

1 Introduction

Smart phones have recently been gaining global attention. The representative aspect that distinguishes smart phones from other mobile phones is that users can install applications("apps" hereinafter) on their smart phones[1]. Depending on the smart phone model, there are tens of thousands of apps available, and hundreds of new apps are launched every day. Users can use Internet banking services, look up bus or subway information simply by installing the related apps. The App Store for iPhones is a closed market in which the provision and distribution of apps is controlled by Apple. On the other hand, Google's Android Market is an open market with minimal controls, and users and developers are responsible for verification of malicious programs and post-installation management. The high financial value of personal information has led to the growth of personal information theft. On this backdrop, if the users should have responsibility for making the verification of malicious programs, it will only enable

* 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(2011-0026023).

** The Corresponding author.

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 353–360, 2011.

the continued growth and propagation of malicious codes designed for personal information theft.

Lookout, a US mobile security service provider, recently published a report on mobile malware and security trends, which reported that the number of Android apps infected with malware increased from 80 in January 2011 to over 400 in June 2011[2]. It also said that the technologies for malware distribution are continuously evolving. Developers of malware frequently upload "clean" apps to the Android Market. They wait until the apps gain popularity by receiving positive reviews, and then distribute malicious codes through updates.

Recently, the IDs, passwords and resident registration numbers of 35 million subscribers to Korea's leading social network services, Nate On and Cyworld, were leaked[3]. A zombie PC was inspected, and it was found that malicious codes were installed while an app was being developed and downloaded. The incident occurred during the course of PC application updates, but it was attributable to an individual user's failure to verify the reliability of apps and app developers. The same incident may occur with smart phones. With smart phone apps showing explosive growth, it is expected that personal information thefts via smart phones caused by malicious codes will grow at the same pace unless method are introduced to ensure the security and reliability of apps.

To this end, this paper examines the problems with the existing smart phone app verification and distribution procedure, and suggests an app distribution method that offers improved reliability and security using the code signing technology interconnected with PKI. The organization of this paper are as follows: First, the security threats against smart phone apps which have recently been discussed is highlighted in Section 2, and Section 3 featured the app verification methods of major app stores. In Section 4, we proposed a novel app verification model that offers improved reliability and security. Section 5 analyzed the scheme suggested in Section 4, and the final Section 5 provides a conclusion.

2 Security Threats of Smart Phone Apps

Hackers may intercept or fabricate data transmitted via the mobile network, or perform DoS attacks to compromise the availability of smart phones. A criminal may steal a smart phone to acquire material information stored on the phone, or to harm the owner of the phone by illegally using the phone[7]. As mentioned above, the security threats to smart phones exist at various locations, including the mobile network, app servers and the smart phones themselves. This paragraph aims to list security threats caused by the use of malicious apps.

First, when apps infected by malicious codes are installed and used on smart phones, the information stored on the phone may be fabricated or sent to hackers. According to an analysis of approximately 300,000 apps conducted by LookOut the possibility of excessive personal information disclosures, including information related to the user's Location, Call history, Text MSG, Email, Contacts was verified[4]. If apps enabling wiretapping are installed, the contents of a user's voice communication and text messages may be disclosed. Another security threat caused by the use of malicious apps includes the possibility that

a smart phone may be used for DDoS attacks or serve as a SPAM sender, causing unnecessary charges for the user[5]. If apps infected by malicious codes cause excessive battery consumption, the availability of the smart phone may be compromised[6].

3 Current App Verification Methods

3.1 Symbian

Symbian verifies applications through a verification organization operated by Symbian to gain access to specific APIs in advance, and distributes the verified apps with electronic signatures. However, when the apps need the basic capability to be distributed, developers may use the self-signed certificate and immediately distribute the apps without further tests by the verification organization[8]. There is a possibility that developers of malicious codes may use the self-signed certificate at their discretion to distribute malicious apps.

3.2 Apple

The method of Apple's app use can be defined as the closed type. In other words, all apps provided to iPhones are verified and distributed under the control of Apple. [9].The procedure for registering apps to the App Store is as follows.

1.To upload apps to iPhones or iPods, test the apps and register them in the App Store, we must complete the developer registration procedure at the website of Apple.

2. After registering with the App Store as a developer, we may download the SDK from the Apple website. To do this, we must register first to use the SDK. The information required for registration are certificates for app signing, devices to be used for development, App ID used for information sharing among developed programs, and distribution policies.

3. When apps are developed, a developer may register the app to iTunes Connect.

4. After the registration of apps, Apple will assess the apps. During the course of this assessment, Apple determines whether the apps include malicious codes and errors causing problems. As there are no clear assessment criteria available, some apps are rejected. Some users "jailbreak" their phones in order to use unauthorized apps.

3.3 Android

The Android Market is an open app store and uses the term "Market" instead of "Store." A market is a venue where suppliers can meet users, and for this purpose, Google discloses its SDK to anyone, also showing a difference from Apple's approach[10]. While the App Store requires several stages for registration of apps, including preliminary assessment, the Android Market only requires the process of user registration[11].

1. User registration can be done at the Android website. All the items a user is required to enter are the name of the developer, the e-mail address, the website address and the contact number.
2. For the development of Android-based apps, a developer needs to have the Java SDK. As the Java SDK is open, a developer can use it without developer registration.

This paper aims to focus on the app distribution method of the Android Market. The Market enables the free distribution of apps, but it has the inherent risk of the distribution of malicious codes. In this regard, this paper will suggest an app distribution method with reliability for the Android Market by interconnection with PKI, in order to add reliability to the openness of the app market.

4 Proposed Code-Signing Scheme

Before a detailed description of the scheme, we'd like to define the requirements and assumptions.

4.1 Security Requirements and Assumptions

The security requirements of the proposed scheme in this paper are as follows.

(1) Integrity of Apps : Apps transmitted from the app market to the smart phones must not be forged or modified by malicious hackers. In other words, the apps shall not be changed during the course of storage at the market or transmission to smart phones.

(2) App Developer Authentication : One of the representative aspects of the Android App Market is that anyone can develop and distribute apps. In other words, the Market is open for hackers. Therefore, users must be able to verify app developers and their status.

(3) Compatibility of Services : The suggested scheme shall be compatible with existing services, and provide easy of use. It shall also be scalable so that it may be used at various app markets.

In order to meet these requirements, the assumptions of this scheme are as follows:

- The App Market has the function of certificate issuance for code signing of developers. The Certificate of the App Market shall be issued by the nation's root certificate agency or the accredited certificate agency.
- Smart phones must have the code signing verification function. The functions to verify digital signatures using X.509 Certificate, including code signing value verification and certificate verification, shall be pre-installed in the smart phones.
- The App Market has a hash server stores hash values of app source code. The hash value is generated in App Market after it verifies the source code. They are used for verification whether the source code is modified or not after validation of App market.

The full process of the proposed scheme can be shown in Fig.1.

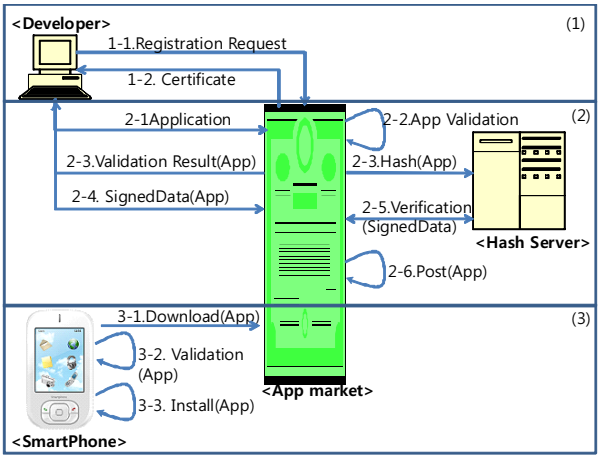


Fig. 1. App validation and Distribution Process using Code Sining

4.2 Step1. Developer Registration

1. A developer applies for developer registration at the market to register the app. A developer submits identification information including name and contact number to the market, and the market issues the app code signing certificate after verifying the submitted information. For the identity proofing of a developer, the market can require some personal information like as social security number, e-mail address, etc., and it should check the identity of a developer using some reliable method.
2. If the identity of a developer is confirmed, the market can issue the code signing certificate for developer. The certificate for a code singing has the profile of an X.509 certificate and $ExtendedKeyUsage$ field in *extensions* should be *codesigning* to indicate for code singing.[12].

4.3 Step2. Development and Distribution of Apps

This step is the procedure for verification of the developed software using code singing scheme.

1. When the developer finished a development of software, he or she should send it to the market.
2. The market must check the vulnerability of the app in advance of its actual release. The market shall thoroughly check vulnerabilities of app source codes and the existence of malicious codes.

3. After completing the verification of app source codes,the market generate hash value of app source and save it in the hash server in the market system. This hash value will be used to check whether the source code of the app is changed or not after validation. And the market sent the validation result of the app the developer.
4. If the validation of the app is passed, a developer generates code signing of the app. It contains *Length*, and *SignatureBlock*. *Length* field tells the length of the code signing and *SignatureBloack* is organized according to PKCS7 as follows[13].

```
SignedData ::= SEQUENCE {
   version Version,
   digestAlgorithm DigestAlgorithmIdentifiers,
   contentInfo ContentInfo,
   certificates [0]    IMPLICIT
           ExtendedCertificatesAndCertificates OPTIONAL,
   crls    [1] IMPLICIT CertificateRevocationLists OPTIONAL,
   signerInfos SignerInfos }
```

   digestAlgorithm is the hash algorithm for hash of app source code like as SHA-256. *contentInfo* is the signed content.It can be generated using the developer's private key. That means the developer generates public key encryption value using his/her private key on the hash value of app. It defines as follows.
   $contentInfo = E_{PR_{developer}}(Hash(App))$
   In here E means a public key encryption and $PR_{developer}$ is the private key of a developer. And $Hash(App)$ is the hash value of app using *digestAlgorithm*. And *certificates* is a set of certificates. It is sufficient to contain chains from a "root" or "top-level certification authority" to the signers in the *signerInfos* field.
   crls is a set of certificate-revocation lists. *signerInfos* can contain a collection of per-signer information.
5. After receiving the signed data from the developer, the App market should verify the integrity of App. To do this, it can use the hash server, that has the hash values of App_i.
   $Hash_{App_i} \; ? = D_{PU_{developer}}(contentInfo)$
   D means a decryption algorithm using a developer's public key($PU_{developer}$) and *contentInfo* is included in *SignedData* sent from a developer.
6. If the hash value of the app software and the hash value in the hash server are same, the market distributes the App on the market server.

4.4 Step3. App Verification and Installation in the Smartphone

1. The user downloads a code-signed app from the app market to his smartphone.

2. A smart phone extracts public keys from the app distributor certificate, decodes the code signature values into key values, and compares the app hash values with the decoded values to verify the integrity of apps. In addition, a user may check whether the app market has issued the app distributor certificate, and verify whether the app distributor is a developer that is authorized by the market through the app developer certificate verification procedure.
3. The user can install the app in the smartphone.

The apps verified through the above three steps are installed and used on smart phones. The upgrades of installed apps may skip step 1.

5 Analysis

This paragraph aims to compare the model suggested in this paper with the existing app verification model, and verify whether the suggested model satisfies the security requirements defined in chapter 4.1.

Table 1. Comparison between the existing scheme and the proposed scheme

Items	Android	Proposed scheme
Verification of Source Codes	Lack of verification procedure	Reinforced verification functions at the market
Verification of Developers	Not	Supported through the developer certificate verification
Code Signing	Self-signed	Signatures through the reliable certificate
Risk of App Modification	High	Low

As seen in the table above, the suggested model enables verification of fabrication or modification of apps using code signing technology. In addition, the model enables the verification of developer identification through verification of the developer certificates used for app code signing.

As for the security requirements defined in chapter 4.1, our proposed scheme satisfies them as follows. The first security requirement is integrity of apps. Modification of apps for malicious purposes must be prevented. The code signing technology in our paper disables the verification of signatures when codes are modified, enabling the integrity of codes to be guaranteed. The second security requirement is app developer verification, which is enabled through the verification of certificates issued to developers by the market. A developer shall submit proper identification information to the market for certificate issuance, and the market shall issue the certificate after verifying the submitted information. A user checks the chain of certificates, including developer certificates to verify

developer certificates. The final security requirement is compatibility of services. The code signing profile used in the model suggested in this paper is already being used globally through X.509 e-signature certification, and is also used for ActiveX in Internet Explorer. In addition, the functions that can be used to internally generate and verify digital signatures at smart phones are applied to various applications.

6 Conclusion

This paper suggested an app verification system using code signing technology to improve the security and reliability of existing app verification systems. The e-signing technology using X.509 certificates supports the integrity of signatures and the signor authentication function. The best solution to the recent emergence of security threats to smart phones concentrated on app stores is to improve the reliability of apps. The vulnerability of app stores can be eliminated by determining whether apps include malicious codes, and whether apps have been developed by malicious hackers. Accordingly, the model suggested in this paper is expected to improve the security of app stores.

References

1. Ballagas, R., et al.: The smart phone: a ubiquitous input device. IEEE Pervasive Computing (2006)
2. Mobile Threats, https://www.mylookout.com
3. The Korea Times, Personal info of 35 mil. Nate, Cyworld users feared leaked (July 28, 2011)
4. Raor, L.: Lookout Identifies Which iPhone And Android Apps Want Your Sensitive Data (July 27, 2010), http://techcrunch.com
5. Guo, C., Wang, H.J., Zhu, W.: Smart-Phone Attacks and Defenses. In: HotNets III (November 2004)
6. Mulliner, C.R.: Security of Smart Phone, Master's Thesis of University of California (June 2006)
7. Jeon, W., Kim, J., Lee, Y., Won, D.: A Practical Analysis of Smartphone Security. In: Smith, M.J., Salvendy, G. (eds.) HCII 2011, Part I. LNCS, vol. 6771, pp. 311–320. Springer, Heidelberg (2011)
8. http://developer.symbian.org
9. http://seriot.ch/resources/talkspapers/iPhonePrivacy.pdf
10. Shabtai, A., Fledel, Y., Kanonov, U., Elovici, Y., Dolev, S.: Google Android: A State of the Art Review of Security Mechanisms, arXiv 2009 (November 2009)
11. Shabtai, A., Fledel, Y., Kanonov, U., Elovici, Y., Dolev, S.: Google Android: A State-of-the-Art Review of Security Mechanisms, http://arxiv.org/ftp/arxiv/papers/0912/0912.5101.pdf
12. Housley, R., et al.: Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile (April 2002)
13. An RSA Laboratories Technical note, PKCS7 : Cryptographic Message Syntax Standard (November 1, 1993)

A Variant of Schnorr Identity-Based Identification Scheme with Tight Reduction

Syh-Yuan Tan[1], Swee-Huay Heng[2], Raphael C.-W. Phan[3], and Bok-Min Goi[4]

[1] Faculty of Information Communication and Technology, Tunku Abdul Rahman University
Perak, Malaysia
tsyuan@utar.edu.my
[2] Faculty of Information Science and Technology, Multimedia University
Melaka, Malaysia
shheng@mmu.edu.my
[3] Department of Electronic and Electrical Engineering, Loughborough University
Leicestershire, UK
r.phan@lboro.ac.uk
[4] Faculty of Engineering and Science, Tunku Abdul Rahman University
Kuala Lumpur, Malaysia
goibm@utar.edu.my

Abstract. Using the rewinding technique from Reset Lemma, Schnorr identity-based identification (IBI) can be proven secure against impersonation under passive attack, and active and concurrent attacks if the discrete logarithm problem and one-more discrete logarithm problem are hard in the random oracle model respectively. However, its security reduction is not tight. In this paper, we propose a variant of Schnorr IBI scheme and provide a direct proof with tight security reduction. In particular, we show that with only three additional elements in the system parameters, the proposed scheme can be proven secure against impersonation under passive attack, and active and concurrent attacks if the decisional Diffie-Hellman problem is hard in the random oracle model. This proving technique may also be applied on other IBI schemes.

Keywords: Identity-based, identitification, tight reduction.

1 Introduction

An identification scheme assures one party (through acquisition of corroborative evidence) of both the identity of a second party involved, and that the second party was active at the time the evidence was created or acquired [12]. Informally speaking, an identification protocol is an interactive process that enables a prover holding a secret key to identify himself to a verifier holding the corresponding public key.

On the other hand, the concept of identity-based (ID-based) cryptography was introduced by Shamir in 1984 [15] where the public key is the user's public identity string (e.g. name, ID number, email, etc.). A trusted third party, namely, the private key generator (PKG) is required to generate private key for every user based on their public identity. There was no rigorous definition as well as security proof until the independent works of Kurosawa and Heng [9] and Bellare et al. [3] in 2004. Kurosawa and

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 361–370, 2011.

Heng proposed transforming certain class of standard digital signature schemes to an IBI scheme, while Bellare et al. concerned about transforming standard identification schemes to IBI. Since then, some other identity-based identification (IBI) schemes were published [10, 11, 6, 17, 16, 13].

Although the Schnorr IBI can be obtained easily by applying the transformation frameworks on the Schnorr standard signature scheme, the resulted security proof is not tight. The generality of the transformation frameworks cannot provide scheme-dependent optimisation as opposed to direct proof which will provide tighter security reductions. A security reduction is said to be tight if the probability of breaking an IBI scheme is close (i.e. ≈ 1) to the probability of solving the underlying mathematical hard problem. If the security reduction is tight, it indicates that the IBI scheme is almost as secure as the underlying mathematical hard problem. A non-tight security reduction needs a larger key size k in order to achieve the same level of security. For instance, let the probability of breaking an IBI scheme be $\epsilon_{IBI} \approx \sqrt{\epsilon}$ where ϵ is the probability of solving a hard problem such as discrete logarithm (DLOG) problem. If $\epsilon = 2^{-80}$ when $k = 160$, $\epsilon_{IBI} \approx \sqrt{2^{-80}} \approx 2^{-40}$ which is not acceptable. In order to achieve a desired security level, i.e. $\epsilon_{IBI} \approx 2^{-80}$, the value of k needs to be increased so that ϵ decreases until $\epsilon_{IBI} \approx \sqrt{2^{-160}} \approx 2^{-80}$.

In year 2007, Goh et al. [8, 7] presented two novel techniques to prove the security of standard signature schemes. The authors showed that with only an additional public key and private key, any DLOG-based standard signature scheme can be proven as secure as the decisional Diffie-Hellman (DDH) problem in the random oracle model (ROM). They also exploited the collision-resistant property of cryptographic hash function to prove the security of any RSA-based standard signature scheme with tight security reduction. Combining the Goh et al. proving technique and the Kurosawa-Heng transformation [9], one can trivially obtain from the Goh et al. standard signature scheme an IBI which is secure against impersonation under passive attack in the random oracle model. However, the security against active and concurrent attacks is not known for the transformed scheme.

In the same year of Goh et al.'s work, Arita and Kawashima [1] proposed a variant of Schnorr standard identification scheme which is secure against impersonation under passive attack based on the DLOG assumption and knowledge-of-exponent (KEA1) without random oracle. They also proved that the scheme is secure against impersonation under active and concurrent attacks based on the one-more DLOG (OMDL) assumption and KEA1 without random oracle. These proofs can achieve tight security reduction by eliminating the need of rewinding the adversary as in Reset Lemma [4] where the simulator can open the commitment from two items which are the simulator transcript and the non-black-box extractor of KEA1. According to the Bellare et al. transformation [3], similar to the Okamoto standard identification scheme [3], the Schnorr standard identification [14] and its variant is not captured by any category of the convertible standard identification scheme. Thus, it indicates that even though we can transform the Arita and Kawashiwa's standard identification scheme to an IBI, we still need to provide a direct proof for it.

The question of whether a transformed IBI from either a standard signature or an identification scheme is secure against active and concurrent attacks has been answered

in the Yang et al. IBI framework [17]. The framework shows that if there exists trapdoor strong-one-more relation and witness dualism proof of knowledge in an IBI scheme, then the IBI scheme is secure against active and concurrent attacks. This was further illustrated by constructing an IBI scheme using the Goh et al. signature [8,7] scheme as the building block. However, the resulted security reduction is also not tight as rewinding of adversary is needed in the security proof to extract a witness.

In this paper, inspired by the technique in [8,7], we propose a variant of Schnorr IBI which requires only three additional elements in the system parameters in order to achieve a tight security reduction. Besides, the proposed scheme can be proven secure against impersonation under passive, active and concurrent attacks based on the DDH assumption in the ROM. This technique may be applied on other IBI schemes as well.

This paper is organised as follows. In Section 2, we provide some preliminaries and definitions. In Section 3, we show the construction of the proposed Schnorr IBI variant and followed by the security analysis in Section 4. Finally, we draw the conclusion in Section 5.

2 Preliminaries

2.1 Decisional Diffie-Hellman (DDH) Assumption [7]

Definition 1. *Let $\mathbb{G}$ be a cyclic group with prime order q. Let g be a generator of $\mathbb{G}$. The decisional Diffie-Hellman assumption states that there exists no polynomial time algorithm M which is able to (t, ϵ_{DDH})-distinguishes a DH-tuple (g, g^x, g^y, g^{xy}) for random $x, y \in \mathbb{Z}_q$ from a random tuple (g, g^x, g^y, g^z) for random $x, y \in \mathbb{Z}_q$ with non-negligible probability such that:*

$$|\Pr[x, y \leftarrow \mathbb{Z}_q : M(g, g^x, g^y, g^{xy}) = true] - \Pr[x, y, z \leftarrow \mathbb{Z}_q : M(g, g^x, g^y, g^z) = true]| \geq \epsilon_{DDH}$$

2.2 Identity-Based Identification (IBI) Scheme [9,3]

Definition 2. *An IBI scheme $\mathcal{IBI} = (\mathcal{S}, \mathcal{E}, \mathcal{P}, \mathcal{V})$ consists of four probabilistic polynomial time algorithms, namely, setup, extract, proving and verification (identification protocol).*

1. ***Setup* ($\mathcal{S}$).** *$\mathcal{S}$ takes as input the security parameter 1^k. It generates the master public key mpk and the master secret key msk. The master public key will be publicly known while the master secret key will be known to the PKG only.*
2. ***Extract* ($\mathcal{E}$).** *PKG takes as input mpk, msk and a public identity ID. It runs $\mathcal{E}$ to extract a user secret key usk.*
3. ***Identification Protocol* ($\mathcal{P}$ *and* $\mathcal{V}$).** *$\mathcal{P}$ receives as input (mpk,usk,ID) and $\mathcal{V}$ receives as input (mpk, ID) and usk is the user secret key corresponding to the public identity ID. $\mathcal{P}$ and $\mathcal{V}$ will run an interactive protocol which consists of the following steps:*

(a) ***commitment:*** *$\mathcal{P}$ sends a commitment CMT to $\mathcal{V}$*
(b) ***challenge:*** *$\mathcal{V}$ sends a challenge CH to $\mathcal{P}$*
(c) ***response:*** *$\mathcal{P}$ sends a response RSP to $\mathcal{V}$*

Finally, $\mathcal{V}$ outputs a boolean decision 1 (accept) or 0 (reject) based on RSP. A legitimate $\mathcal{P}$ should always be accepted.

2.3 Security Model

The impersonation attack game on an IBI scheme between an impersonator I and a challenger is described as a two-phased game [9,3] as follows:

1. **Setup.** The challenger takes as input 1^k and runs the setup algorithm $\mathcal{S}$. It gives I the resulting master public key mpk and keeps the master secret key msk to itself.
2. **Phase 1**
   (a) I issues some extract queries on $ID_1, ID_2, \ldots$. The challenger responds by running the extract algorithm $\mathcal{E}$ to generate the user secret key usk_i corresponding to the public identity ID_i. It returns usk_i to I.
   (b) I issues some transcript queries for passive attack or some identification queries on ID_j for active/concurrent attack.
   (c) The queries in step (a) and step (b) above can be interleaved and asked adaptively. Without loss of generality, we may assume that I will not query the same ID_i that has been issued in the extract queries, in the transcript queries or identification queries again.
3. **Phase 2**
   (a) I plays the role as a cheating prover (impersonation attempt on the prover holding the challenged identity ID' such that ID' is not issued an extract query before), trying to convince the verifier.
   (b) I can still issue some extract queries as well as transcript queries or identification queries as in Phase 1.

Definition 3. *We say that $\mathcal{IBI}$ is (t, q_e, ϵ_{IBI})-secure under passive (active and concurrent) attack if for any passive (active and concurrent) impersonator I who runs in time t,*

$$\Pr[I \textit{ can impersonate}] \leq \epsilon_{IBI}$$

where I can make at most q_e extract queries and ϵ_{IBI} is the advantage of breaking an IBI scheme.

3 Schnorr IBI

We first give the construction of the Schnorr IBI scheme which is a transformation from standard Schnorr signature [14] scheme by using the Kurosawa and Heng transformation [9]:

Setup. On input 1^k, generate two large primes p and q such that $q|(p-1)$. Choose $x \xleftarrow{R} \mathbb{Z}_q$ to compute $y_1 = g^{-x}$ where $g \xleftarrow{R} \mathbb{G}$. Let the master public key be $mpk = (p, q, g, y_1, H)$ and the master secret key be $msk = x$ where $H : \{0,1\}^* \times \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{Z}_q$.

Extract. Let ID be the public identity. Select a value $t \xleftarrow{R} \mathbb{Z}_q$, compute $A = g^t$ and $s = t + x\alpha$ where $\alpha = H(ID, A, y_1)$. Return the user secret key as $usk = (\alpha, s)$.

Identification Protocol

1. $\mathcal{P}$ first computes $A = g^s y_1^{\alpha}$. $\mathcal{P}$ next chooses $r \xleftarrow{R} \mathbb{Z}_q$, computes $X = g^r$ and sends (A, X) to $\mathcal{V}$.
2. $\mathcal{V}$ chooses $c \xleftarrow{R} \mathbb{Z}_q$ and sends c to $\mathcal{P}$.
3. $\mathcal{P}$ computes $y = r + cs$ and sends y to $\mathcal{V}$.
4. $\mathcal{V}$ accepts if and only if $g^y = X \cdot (A/y_1^{\alpha})^c$ where $\alpha = H(ID, A, y_1)$.

Correctness:

$$\begin{aligned} X(A/y_1^{\alpha})^c &= g^r(g^s g^{-x\alpha}/g^{-x\alpha})^c \\ &= g^r(g^s)^c \\ &= g^{r+cs} \\ &= g^y \end{aligned}$$

If the equality holds, output **1 (accept)**, else output **0 (reject)**.

3.1 A Variant of Schnorr IBI

The construction is similar to the technique used in [8,7] which aims to eliminate the use of forking lemma in proving the security of signature schemes. The differences of this variant from the original scheme in Section 3 are: 1) Two additional elements of $\mathbb{G}$ are included in mpk 2) One additional element of $\mathbb{G}$ is included in msk. We now show the variant of Schnorr IBI which takes three additional elements in the system parameters as follows:

Setup. On input 1^k, generate two large primes p and q such that $q|(p-1)$. Choose $x \xleftarrow{R} \mathbb{Z}_q$ to compute , $y_1 = g^{-x}, y_2 = h^{-x}$ where $g, h \xleftarrow{R} \mathbb{G}$. Let the master public key be $mpk = (p, q, g, h, y_1, y_2, H)$ and the master secret key be $msk = x$ where $H : \{0,1\}^* \times \mathbb{G} \times \mathbb{G} \times \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{Z}_q$.

Extract. Let ID be the public identity. Select a value $t \xleftarrow{R} \mathbb{Z}_q$, compute $A = g^t, B = h^t$ and $s = t + x\alpha$ where $\alpha = H(ID, A, B, y_1, y_2)$. Return the user secret key as $usk = (\alpha, s)$.

Identification Protocol

1. $\mathcal{P}$ first computes $A = g^s y_1^{\alpha}$ and $B = h^s y_2^{\alpha}$. $\mathcal{P}$ next chooses $r \xleftarrow{R} \mathbb{Z}_q$, computes $X = g^r$ and sends (A, B, X) to $\mathcal{V}$.
2. $\mathcal{V}$ chooses $c \xleftarrow{R} \mathbb{Z}_q$ and sends c to $\mathcal{P}$.
3. $\mathcal{P}$ computes $y = r + cs$ and sends y to $\mathcal{V}$.
4. $\mathcal{V}$ accepts if and only if $g^y = X \cdot (A/y_1^{\alpha})^c$ where $\alpha = H(ID, A, B, y_1, y_2)$.

Correctness:

$$\begin{aligned} X(A/y_1^{\alpha})^c &= g^r(g^s g^{-x\alpha}/g^{-x\alpha})^c \\ &= g^r(g^s)^c \\ &= g^{r+cs} \\ &= g^y \end{aligned}$$

If the equality holds, output **1 (accept)**, else output **0 (reject)**.

Notice that the value B is not involved in the verification of prover's response and this is due to the reason that the value B is bonded with usk by a secure hash function H. Recall that the normal way of proving the security of an IBI is by exploiting the proof of knowledge (POK) of a discrete logarithm in the **Identification Protocol** using Reset Lemma [4] (which resulted in non-tight reduction). The trick of our proof is that we are getting the prover I to prove that y (and subsequently s) contains the information of the discrete logarithm of y_1 and y_2, instead of proving the knowledge of the discrete logarithm and this does not require the help of B during the interaction. Besides, since the hash function H in **Extract** binds B with the scheme parameters, there is no way an adversary makes the verifier outputs **1** with an altered B value except with negligible probability (e.g. collision in H). We can save some complexity by not adding the value B and the protocol is still as secure as the original Schnorr IBI scheme.

4 Security Analysis

4.1 Security against Impersonation under Passive Attack

Theorem 1. *The above IBI scheme is (t, q_e, ϵ_{IBI})-secure against impersonation under passive attack in the random oracle model if the decisional Diffie-Hellman assumption (DDH) holds such that:*

$$\begin{aligned} t &\leq t' - 2.4(q_e + 1)t_{exp} \\ \epsilon_{IBI} &\geq \epsilon_{DDH} + 2(q_e + 1)q^{-1} \end{aligned}$$

where q_e is the total extract queries that are queried by an impersonator I and assuming a two-exponent multi-exponentiation takes time $1.2t_{exp}$.

Proof. Given the tuple (g, h, y_1, y_2) as input, we construct an algorithm M running in time t' that can determine whether the given tuple is a DH tuple or not with the help of the impersonator I.

Phase 1
Setup. M sets $mpk = (p, q, g, h, y_1, y_2, H)$ where H is modeled as a random oracle which takes five inputs such that $H : \{0,1\}^* \times \mathbb{G} \times \mathbb{G} \times \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{Z}_q$.

Extract Query. For an extract query on ID queried by I, M selects $s, \alpha \xleftarrow{R} \mathbb{Z}_q$ to compute $A = g^s y_1^{\alpha}$ and $B = h^s y_2^{\alpha}$. M sets $\alpha = H(ID, A, B, y_1, y_2)$ and returns

$usk_{ID} = (s, \alpha)$ to I. Notice that two additional two-exponent multi-exponentiations are required and this lead to the additional complexity of $2.4t_{exp}$ per usk_{ID} extraction [7,11].

Transcript Query. For a transcript query on ID queried by I, M first checks if ID has been queried an extract query before. If yes, M uses the existing usk_{ID} to return a valid transcript for I. If no, M runs extract query algorithm to generate usk_{ID} and returns a valid transcript for I.

Phase 2
I pretends to be a valid prover of an identity ID^* which has not been queried an extract query before and runs the identification protocol with M. At the end of the identification protocol, M obtains a transcript (A, B, X, c, y) on ID^*. If the transcript is not valid, M aborts and it fails in the security game. Else if the transcript is valid, M can determine whether the given tuple is a DH tuple and wins in the security game with the probability as follows:

$$\begin{aligned}
&\Pr[M \text{ wins}] \\
&= \Pr[M \text{ accepts prover}] - \Pr[M \text{ aborts if DH tuple}] - \Pr[M \text{ not aborts if random tuple}] \\
&\leq \epsilon_{IBI} - \Pr[M \text{ aborts if DH tuple}] - \Pr[M \text{ not aborts if random tuple}]
\end{aligned}$$

We first examine the probability that M aborts if the tuple is a DH tuple. If it is a DH tuple, M simulates the IBI perfectly except with the negligible probability q^{-l} on the collision of the random oracle H when answering I's extract queries for $q_e + 1$ times.

Secondly, if the tuple is a random tuple, M does not abort with probability q^{-1} each time answering I's extract queries for q_e times and does not abort with probability q^{-1} when I made a call to the random oracle H in Phase 2. The probability of not abort depends on the value s such that there is at most one possible value of α for which there exists an s satisfying $A = g^s y_1^\alpha$ and $B = h^s y_2^\alpha$.

Combine the probabilities together, we get:

$$\begin{aligned}
\Pr[M \text{ wins}] &\leq \epsilon_{IBI} - (q_e + 1)q^{-l} - (q_e + 1)q^{-1} \\
\epsilon_{DDH} &\leq \epsilon_{IBI} - 2(q_e + 1)q^{-1}
\end{aligned}$$

as required. □

4.2 Security against Impersonation under Active and Concurrent Attacks

Theorem 2. *The above IBI scheme is (t, ϵ_{IBI})-secure against impersonation under active and concurrent attack in the random oracle model if the decisional Diffie-Hellman assumption (DDH) holds such that:*

$$\begin{aligned}
t &\leq t' - 2.4(q_e + 1)t_{exp} \\
\epsilon_{IBI} &\geq \epsilon_{DDH} + 2(q_e + 1)q^{-1}
\end{aligned}$$

where q_e is the total extract queries that are queried by an impersonator I and assuming a two-exponent multi-exponentiation takes time $1.2t_{exp}$.

Proof. Given the tuple (g, h, y_1, y_2) as input, we construct an algorithm M running in time t' that can determine whether the given tuple is a DH tuple or not with the help of the impersonator I.

Phase 1
Setup. Same as the proof in Section 4.1.

Extract Query. Same as the proof in Section 4.1.

Identification Query. For an identification query on ID by I, M first checks if ID has been queried an extract query before. If yes, starting with the clone $m = 1$, M uses the usk_{ID} to return a valid transcript for I. If no, M runs extract query algorithm to generate usk_{ID} and plays the role of prover as in the identification protocol of Section 3.1 starting with the clone $m = 1$:

1. M first computes $A = g^s y_1^{\alpha}$ and $B = h^s y_2^{\alpha}$. M next chooses $r \xleftarrow{R} \mathbb{Z}_q$, computes $X = g^r$ and sends (A, B, X) to I.
2. I sends $c \xleftarrow{R} \mathbb{Z}_q$ to M.
3. M computes $y = r + cs$ and sends y to I.
4. M increases m for 1.

Phase 2
I pretends to be a valid prover of an identity ID^* which has not been queried an extract query before and runs the identification protocol with M. At the end of the identification protocol, M obtains a transcript (A, B, X, c, y) on ID^*. If the transcript is not valid, M aborts and it fails in the security game. Else if the transcript is valid, M can determine whether the given tuple is a DH tuple and wins in the security game with the same probability as the proof in Section 4.1. If there remains any unanswered query of other clone m, M reacts as in the identification query algorithm.

Up to date, the proof of security against active and concurrent attacks for IBI must involve one-more problems and Reset Lemma but we manage to avoid that. Similar to the work of [8,7], this shows that a tighter security reduction can be achieved by basing the security of a scheme on a stronger assumption. □

4.3 Reset Attacks

The resettable attacks [2] grant the adversary the power of resetting the identification protocol to any state it wants. It is obvious that the proposed Schnorr IBI variant which is applying the zero-knowledge proof of knowledge protocol cannot resist such attack. When the adversary pretends to be a verifier is running an identification protocol with a prover, after receiving the prover's response y_1 on the challenge c_1, it can reset the prover (which is normally a smart card) back to the state where the prover has just sent the commitment A, B, X (by terminating the power supply to smart card). The

adversary will now send a second challenge c_2 to the prover and get the second response y_2 from prover. At the end, the adversary can compute the prover's secret such that $s = (y_2 - y_1)/(c_2 - c_1)$.

However, we are not considering resettable attacks in this paper and we argue that such attack is an implementation issue which can be prevented. For instance, we can program the prover to delete the commitment value r from the memory before sending the response y to the verifier. In this case, whenever the prover is reset, the identification protocol cannot be completed due to the fact that the prover no longer has the knowledge of the commitment value r and will abort during the computation of response (i.e. null value error).

5 Conclusion

We proposed a variant of Schnorr IBI scheme and showed that a tighter security reduction can be achieved by basing the security of the scheme on a stronger assumption. More precisely, we supported the Schnorr IBI variant with a direct security proof which yields a tight security reduction. The scheme is proven secure against impersonation under passive, active and concurrent attacks based on the DDH assumption in the random oracle model.

References

1. Arita, S., Kawashima, N.: An Identification Scheme with Tight Reduction. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. E90-A(9), 1949–1955 (2007)
2. Bellare, M., Fischlin, M., Goldwasser, S., Micali, S.: Identification Protocols Secure against Reset Attacks. In: Pfitzmann, B. (ed.) EUROCRYPT 2001. LNCS, vol. 2045, pp. 495–511. Springer, Heidelberg (2001)
3. Bellare, M., Namprempre, C., Neven, G.: Security Proofs for Identity-Based Identification and Signature Schemes. In: Cachin, C., Camenisch, J.L. (eds.) EUROCRYPT 2004. LNCS, vol. 3027, pp. 268–286. Springer, Heidelberg (2004)
4. Bellare, M., Palacio, A.: GQ and Schnorr Identification Schemes: Proofs of Security against Impersonation under Active and Concurrent Attacks. In: Yung, M. (ed.) CRYPTO 2002. LNCS, vol. 2442, pp. 162–177. Springer, Heidelberg (2002)
5. Bellare, M., Palacio, A.: The Knowledge-of-Exponent Assumptions and 3-Round Zero-Knowledge Protocols. In: Franklin, M. (ed.) CRYPTO 2004. LNCS, vol. 3152, pp. 273–289. Springer, Heidelberg (2004)
6. Chin, J.-J., Heng, S.-H., Goi, B.-M.: An Efficient and Provable Secure Identity-Based Identification Scheme in the Standard Model. In: Mjølsnes, S.F., Mauw, S., Katsikas, S.K. (eds.) EuroPKI 2008. LNCS, vol. 5057, pp. 60–73. Springer, Heidelberg (2008)
7. Goh, E.-J., Jarecki, S., Katz, J., Wang, N.: Efficient signature schemes with tight reductions to the Diffie-Hellman problems. Journal of Cryptology 20(4), 493–514
8. Katz, J., Wang, N.: Efficiency improvements for signature schemes with tight security reductions. In: ACM — CCS 2003, pp. 155–164 (2003)
9. Kurosawa, K., Heng, S.-H.: From Digital Signature to ID-based Identification/Signature. In: Bao, F., Deng, R., Zhou, J. (eds.) PKC 2004. LNCS, vol. 2947, pp. 248–261. Springer, Heidelberg (2004)

10. Kurosawa, K., Heng, S.-H.: Identity-Based Identification Without Random Oracles. 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. 603–613. Springer, Heidelberg (2005)
11. Kurosawa, K., Heng, S.-H.: The Power of Identification Schemes. In: Yung, M., Dodis, Y., Kiayias, A., Malkin, T. (eds.) PKC 2006. LNCS, vol. 3958, pp. 364–377. Springer, Heidelberg (2006)
12. Menezes, A.J., van Oorschot, P.C., Vanstone, S.A.: Handbook of Applied Cryptography. CRC Press (1997)
13. Rückert, M.: Adaptively Secure Identity-Based Identification from Lattices without Random Oracles. In: Garay, J.A., De Prisco, R. (eds.) SCN 2010. LNCS, vol. 6280, pp. 345–362. Springer, Heidelberg (2010)
14. Schnorr, C.-P.: Efficient Identification and Signatures for Smart Cards. In: Brassard, G. (ed.) CRYPTO 1989. LNCS, vol. 435, pp. 239–252. Springer, Heidelberg (1990)
15. Shamir, A.: Identity-Based Cryptosystems and Signature Schemes. In: Blakely, G.R., Chaum, D. (eds.) CRYPTO 1984. LNCS, vol. 196, pp. 47–53. Springer, Heidelberg (1985)
16. Thorncharoensri, P., Susilo, W., Mu, Y.: Identity-Based Identification Scheme Secure against Concurrent-Reset Attacks without Random Oracles. In: Youm, H.Y., Yung, M. (eds.) WISA 2009. LNCS, vol. 5932, pp. 94–108. Springer, Heidelberg (2009)
17. Yang, G., Chen, J., Wong, D.S., Deng, X., Wang, D.: A new framework for the design and analysis of identity-based identification schemes. Theoretical Computer Science 407, 370–388 (2008)

Implementation of Clinical Decision Support System Architecture

Jeong Ah Kim[1], Min Hee Choi[2], and InSook Cho[2]

[1] Computer Education Depart., Kwandong University, 522 NaeKok Dong, KangNung, Korea
[2] Nursing Depart., Inha University, Incheon, Korea
clara@kd.ac.kr, cmhcmh79@lycos.co.kr, insook.cho@inha.ac.kr

Abstract. Quality control in medical is getting very important issue so that the importance of CDS(Clinical Decision Support) System has been increased. Local clinics as well as big hospitals are required to implement the CDS System. But the cost and complexity of CDS system implementation is so high since many different activities including knowledge authoring, software development, and integrating the legacy system are necessary. In this paper, we suggest the CDS system architecture to be sharable and interoperable and evaluate the availability and efficiency of this architecture.

Keywords: architecture, CDSS, interoperability.

1 Introduction

The greatest contribution of medical information technology in the medical field would be performing the duties beyond human limitations instead [1]. While (Clinical Decision Support: hereinafter CDS) system falls under the safety device that can protect the patient safety and medical staffs at the same time, the reason why such safety device is necessary is because our clinical practice is not always an exemplary one. The CDSS related research and development of developed countries including the United States has about 50 year's history. Compared to the fact that most research and development have wide range of application from being rule-based centered on medical information academic circles to various artificial intelligence technology or Machine Learning, the clinical practice and industrial circles prefer a rule-based system [2]. Although our country is also attempting to implement CDS service by data mining method, there are no cases of actual clinical application.

There was a case of defining the standard that a medical information system must possess on a nationwide level from medial perspective and information perspective through EHR Common Core Technology Research and Development Project Team (hereinafter EHR Project Team) from 2005 to 2010 by the Ministry of Health and Welfare.

One objective of EHR Project Team is to implement the method that can define the knowledge to be reused by all hospitals in the field of CDS service and the environment to execute these knowledge independently from the existing information system to present a method that can be easily integrated with the existing system. In

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 371–377, 2011.

this paper, The CDS architecture which is a part of CDS service research results of EHR Project team will be defined, implemented result will be presented and the applicability will be verified.

In chapter 2, the necessity of a new CDS system architecture design and requirements of CDS architecture on a nationwide level will be defined. In chapter 3, the CDS system architecture which is a result of research performed by EHR Project Team will be defined. In chapter 4, the architecture will be evaluated by analyzing the result of implementing architecture at 3 large hospitals. In chapter 5, the method of using study results and future study objectives will be organized.

2 Limitations of Existing CDS System Implementation

2.1 CDS System Overview

In a medical field, a system supporting diagnosis in the clinical practice of specialist is called a CDS system. It is an information system supporting the planning, operation and control function of organization by providing the information at a proper time for the decision makers of hospital[7]. The purpose of developing and using CDSS is to provide support on more precise decision making by minimizing errors from uncertainty through proper data, information and knowledge management that cannot be performed by human beings with the help of a computer system. In terms of medical field, the purposes for developing intelligent DSS or expert system are as follows [8].

(1) To improve the accuracy of clinical diagnosis by integrating data from system oriented, complete and distributed resources. (2) To improve the reliability of clinical decision making by preventing from the influence of cases having similarities that are unidentical and not verified. (3) To improve the cost effect of examination and treatment by placing the balance on inconvenience and time consumption related to the risks and benefits of decisive actions. (4) To enhance understanding on the structure of medical knowledge by relating with the advancement of technology for clarifying the inconsistent and inaccurate knowledge. (5) To enhance understanding on the clinical decision making with the effort for developing more efficient and easy to use system to improve medical training and help understanding on medical science.

2.2 Existing CDS System Architecture

As the CDSS architecture proposed at the early stage was very simple, it was at a level of entering cases manually and providing diagnosis consisted of explanatory note as the one on the premise of stand-alone application method. The SAGE architecture proposed by Stanford University research team in 2006 [3] has proposed a knowledge engine based CDS system architecture that can execute the reusable know ledge in compliance with the medical standard through the connection with VMR(Virtual Medical Record) interface and terminology server. SAGE architecture has a precondition that medical information system is set up at each hospital and each

hospital is using standard terms based on the VMR defined by HL 7 which is an agency defining medical standards. However, many hospitals of our country have defined independent medical information system data models and defines independent terminology, this is a model which is not suitable for domestic conditions considering the mutual manageability of knowledge.

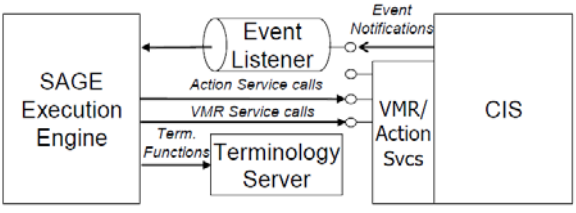


Fig. 1. SAGE System Architecture

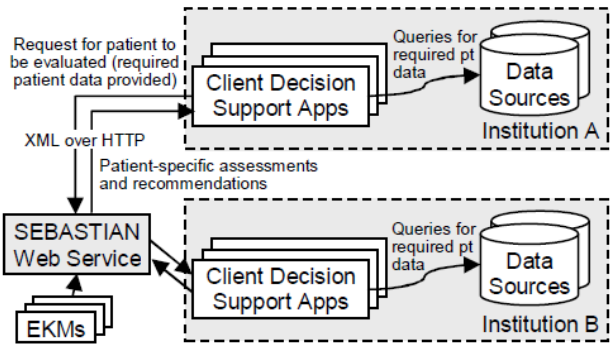


Fig. 2. SABATIAN System Architecture

SABATIAN architecture is a model that has not separated application and knowledge with disadvantages of not being able to reuse by separating knowledge and service or not being able to guarantee the mutual manageability that the SAGE model has.

3 CDS System Architecture Design of EHR Perspective

3.1 Basic Details of Setting Up the Architecture

There are 3 types of most important key requirements for improvement to supplement the existing CDS system architecture.

1) The module that has implemented the knowledge necessary for CDS implementation and CDS service must be separated to secure maintainability and expandability of each.

2) The knowledge and knowledge execution module written by specialists must be integrated easily with various hospital information systems.
3) The changed details of existing system for adding new CDS services must be minimized
4) As the components of CDS and implementation methods are diverse, this must be accommodated.

3.2 CDS System Architecture

In order to set up a core CDSS architecture for implementation, four types of architecture principles have been established in this study [1,6].

First, establish an integrated architecture. Second, establish a component-based adaptable architecture. Third, establish as an architecture than can be implemented. Fourth, establish as a customized architecture.

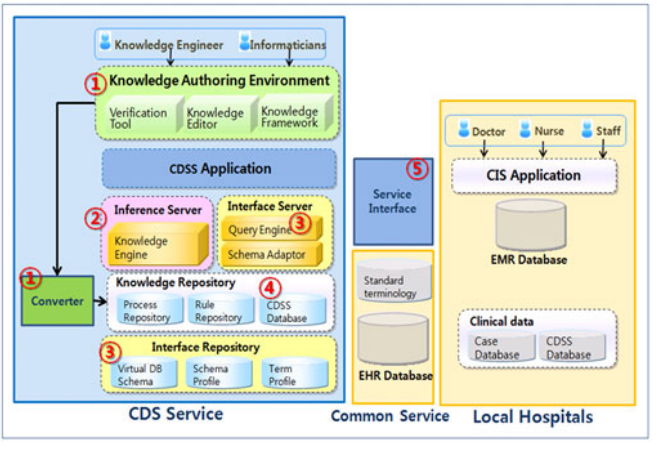


Fig. 3. CDSS architecture in EHR context

From the EHR's perspective, the same knowledge must be used at different hospital environments and must be able to guarantee a flexible integration with the existing application. To make this possible, the application architecture has been defined as five main building blocks (① writing environment module ② knowledge engine module ③ data interface adapter module ④ knowledge storage module ⑤ service interface module).

(1) Writing Environment Module: This is an environment to express knowledge based on rule and process. to convert the written knowledge as a form that computer can understand. It supports test case creator and auto test executor that can support verification which is the core of knowledge writing process to enable improved productivity of knowledge writing and verification.

(2) Knowledge Engine Module: This was organized as rule engine and process engine as engines to execute the clinical knowledge. By doing so, the process based knowledge and all basic rules consisted of if-then can be processed [10].
(3) Data Interface Adapter: From the EHR's perspective, the knowledge cannot be written as a subordinate knowledge of a specific hospital. Therefore, the knowledge must be written based on eVMR(extensible VMR) that has extended VMR(Virtual Medical Record) promoted as HL7 standard to maintain generality. The data interface adapter was implemented from the components necessary for maintaining generality.
(4) Knowledge Storage Module: From the EHR's perspective, the knowledge is an asset that can be shared between various medical institutions. An environment to share the knowledge as an asset was implement by implementing a component to store various types of knowledge by making into meta data, collaborative environment to exchange knowledge verification feedback along with change of knowledge and shape management module.
(5) Service Interface Module: This paper proposes two types of methods. The connection type method is the case of making changes by calling a new service at the existing hospital information system while implementing a screen necessary for CDSS service separately. The integration type method is a method requiring module revision for changing screen, calling CDSS service or processing results of existing hospital information system.

4 How to Apply the Architecture

4.1 Deployment Process

The architecture proposed by this paper and the distribution architecture applying common component to the hospital are shown in Figure 4. In the mandatory electronic record field, the existing information system that can provide patient information and data interface adapter module that can obtain data from the information system are equipped. The newly developed CDSS service module is equipped to the application field and a module that can connected with knowledge engine is equipped. In the engine field, the knowledge execution engine for CDSS service and the data interface adapter module for obtaining additional patient information crated while executing are equipped.

4.2 Verification Results

To verify the architecture proposed in this paper and to verify the effect of CDS operation, the drug interaction CDSS of Ajou University Medical Center and the diagnostic examination CDSS clinical evaluation of Bundang Seoul University Hospital have been performed[1,9]. As seen in Table 1, the knowledge engine is showing a quite stable value and shown as the data interface adapter part occupying most of the time. To verify the architecture made as the study result, the performance and mutual manageability must be evaluated. For the mutual manageability evaluation, one knowledge must be serviced in many hospitals based on same architecture.

However, the first performance verification is necessary for the mutual manageability verification due to the special nature that hospital has. First, the environment same as management department was set up as test environment and performed verification by organizing the same used environment as the situation during management in order to verify the performance.

Table 1. Performance Evaluation (unit: ms)

Practice	# of trans.	Average Response Time		
		Total response time	Engine time	EMRdata query
OutPatient	122,949	197.36 (251,46)	56.00 (29.02)	141.36 (249.17)
InPatient	149,282	626.22 (853.22)	48.02 (17.14)	578.20 (853.80)
ER	51,214	193.98 (247.97)	53.10 (25.16)	140.89 (251.32)

For the system oriented usage and expected effects, the following clinical evaluations are received.

(1) Can be applied to different medical institutions from each other.
(2) Can be serviced targeting many medical institutions as independent form of service
(3) The knowledge extension, subject extension and user extension are flexible without the development of additional system in extending the CDSS service.
(4) In case of diagnostic examination, the effect of architecture gaining the knowledge required while executing the knowledge real-time could be verified through data interface adapter from having the nature of vast quantity of data handed over from the existing mandatory electronic record system. Also, this means being able to secure the mutual manageability of the knowledge written based on general EMR structure through DIA.
(5) Adding, deleting and changing default value on the knowledge is easy under the writing environment.

5 Conclusion

Today, the knowledge based CDS service was started from the approach of expert system imitating human thoughts. For a well developed and verified knowledge infrastructure or model to be widely understood and accommodated in practical work, the support of a tool that can interpret and convert as a form to execute CDS system standardization and standard based knowledge is necessary.

CDS system standardization is able to secure knowledge and raise the productivity by promoting knowledge sharing and reuse between many medical institutions while being able to greatly reduce the additional effort required for localization due to different platform and application environment of each medical institution.

In this study, a CDS system architecture was proposed as a part of standard medical information system installation project of national level. The key components have been implemented to be applied on the actual hospital practice. The applicability of such practice was verified by applying on CDS service implementation based on different knowledge to 3 large hospitals. The CDS system architecture proposed by this paper is a model that can be referred in implementing both the knowledge based and non knowledge based CDS systems.

Therefore, the architecture proposed in this study can be applied while applying the decision making model using data mining or mechanical learning method, or while applying the case based model on the clinical practice. However, the engine supporting the corresponding knowledge expression must be selected and an interface with different module from the engine must be implemented. Also, this can be directly applied to new services extended from u-health and consumer empowerment of individual health records. The verification on mutual manageability of knowledge is necessary through more case applications and additional requirements that can raise effectiveness in the distribution process of knowledge must be identified.

References

1. Cho, I.S., et al.: Technical report of Clinical Decision Support System. In: Elsevia Korea (2010)
2. Ball, M.J.: Back to the future: what have we failed to learn? How does the future look? In: Medinfo 2010. International Medical Informatics Association, CapeTown (2010)
3. Tu, S., Glasgow, J.: SAGE Guideline Model Technical Specification (2006)
4. Kawamoto, K., Lobach, D.F.: Design, implementation, use, and preliminary evaluation of SEBASTIAN. a standards-based Web service for clinical decision support. In: AMIA Annu. Symp. Proc. 2005, 380–384 (2005)
5. Kim, J., Cho, I.S., Kim, Y.: CDSS(Clinical Decision Support System) Architecture in Korea. In: Proceedings of Conference ICHIT (2008)
6. Kim, J., et al.: CDSS architecture in EHR perspectives. In: Proceedings of Software Engineering Summer Workshop (2009)
7. Osheroff, J.A., et al.: Clinical Decision Support Impelementer's Workbook. In: HIMSS (2004)
8. Kawamoto, K.: Service Functional Model Specification - Decision Support Service, DSS (2006)
9. Lee, J.H., Kim, J.A., Cho, I.S., Kim, Y.: Integration of Workflow and Rule Engines for Clinical Decision Support Services. In: MedInfo 2010 Conference, Cafetown (2010)
10. Kim, J., et al.: Translation Protégé Knowledge for Executing Clinical Guidelines. In: Proceedings of Conference Protege 2009 (2009)

ProcessCodi: A Case Study on Social BPM through Integration of SNS, Mind Map, and BPMS

JaeHoon Lee[1], JinYoung Jang[1], and Jeong Ah Kim[2]

[1] uEngine Solutions, 891-37 Daechi-dong, Gangnam-gu, Seoul, Korea
{shamvala,jinyoungj}@gmail.com
[2] Kwandong University, Computer Education Department,
522 Naegok-dong, Gangneung, Korea
clarakja@gmail.com

Abstract. Social BPM is considered as a promising tool to improve the performance of an organization by providing extensible communication tools, informal data handling functions, and knowledge based decision supports. In this study, we developed a Social BPM named ProcessCodi which is integrated by a social network service (SNS), a mind map, and a business process management system (BPMS). Integrating a SNS enables to involve members of the social network into a business process, and a mind map enables to manage the related knowledge as a formal data structure; knowledge tree. A BPMS is a framework to integrate these tools into a process oriented concept. ProcessCodi was implemented in a S/W company and the plausibility are investigated.

Keywords: Social BPM, business process management.

1 Introduction

Recently, integrating business process management (BPM) and social networks has been interested by researchers and practitioners [1][2]. Traditionally, BPM has been considered as an inter-organizational management methodology and focused on standardization and control. On the other hands, social networks are opened, flexible, and hard to be controlled. Therefore, it is required of tools for managing these two different concepts to synergy their effects, therefore social BPM is interested as a promising tool.

Social BPM is basically originated from BPM, thus it has phases of process discovery, modeling, execution, monitoring, and analysis. Recent studies for social BPM are done with collaborative process modeling [3], and social network analysis for discovery [4]. BPM can provide the context of collaboration by leveraging synchronous or asynchronous social networking, and social networking supports and augments the various activities of the BPM application's continuous improvement lifecycle [5]. However social BPM has not been implemented yet, and its benefits and plausibility have not been verified.

In this study, we developed a social BPM named ProcessCodi. ProcessCodi employs an extensible communication tool to interact with people in social networks,

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 378–383, 2011.

a knowledge management tool to handle informal data by mash-ups, and process management tool to involve those resources into a business process. To do so, ProcessCodi was integrated by a social network service (SNS), a mind map, and a business process management system (BPMS). Integrating a SNS enables to involve members of the social network into a business process, and a mind map enables to manage the related knowledge as a formal data structure. A BPMS does a role of a framework to integrate these tools into a process oriented concept.

2 Design Concept: Social Business Process Patterns

Although involving social networks into business is promising and unavoidable, there are two main issues and views on utilizing them. One is "How to raise the value of a company through social networks" and the other is "How to reduce and prevent potential problems from use of social networks". To manage them, they should be discovered, visualized, standardized, defined, planned, and finally improved. These concepts are closed to those of BPM methodology. Moreover, because social networks are bigger structure than an enterprise organization, the patterns of social business processes may be much more complicated.

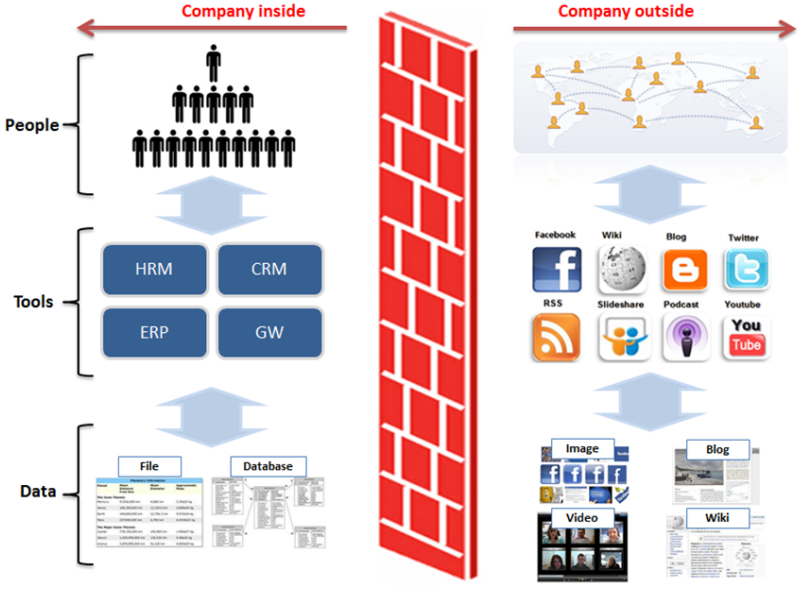


Fig. 1. Business environment: organization structure, IT tools, and data

To implement social network in BPM, we established use cases of social networks named social business process patterns. Social business process patterns refer to a generalized scenario which social networks are utilized by business actors. According to the pattern of a company, utilizing social networks can be either beneficial or

harmful. For example, use of social networks for managing customers is a great trend for recent on-line B2C companies. On the other hands, companies that have severe security policies have to be prudent when to adopt social networks.

As shown in Fig. 1, business environment consists of the inside of an organization which has hierarchical structure, formal data and tasks, enterprise systems, and the outside which has horizontal networks, informal data, and distributed systems. One of the biggest obstacles of utilizing social networks is that these two areas are separated by the wall of an organizational boundary. A social BPM tool should make pipelines into this wall so that workers can retrieve information through them. In addition, it controls the issues on the outer flow which is related to security and the inner flow which is related to information quality or copyright.

Social business process patterns are categorized by the target to be mainly utilized by a business unit. They can be human networks or opened information outside a company. Human networks are used to encourage participation of domain experts or promote a product or a service to potential customers. Patterns using opened information are to capitalize open contents as knowledge of a company through mash-ups. Social business process patterns span over organizations, users, and systems in wide area, thus they have to be implemented by integration of SNS, e-mail, knowledge management tools, BPM.

3 ProcessCodi: Social BPM Implementation

We developed a Social BPM named ProcessCodi, which is a powerful tool to implement social business process patterns. The development strategy of ProcessCodi is as follows.

3.1 Integration with SNS

The basic function of Social BPM is to get workers outside a company involve into the internal business processes. To do so, the BPM should employ the functions to involve the users of social networks. Fig. 2 shows a contact list of ProcessCodi which can include potential users who exist in Facebook for a task. If a user has an expert about your job among Facebook friends, he/she can be involved into the task. At this time, ProcessCodi can control security problems by forcing the participants to get steps of approval.

3.2 Integration with Knowledge Management Tools

The core knowledge of a company is an enterprise strategy. Representatively, a mind-map enables workers to understand, formalize, and share it. ProcessCodi includes a mind-map named OK mind-map so that a user can make a mind-map directly in a BPMS. Also, key nodes of a mind-map can be transformed as a strategy map so that they can be viewed in a navigator, which allows managing knowledge and processes in an integrated environment.

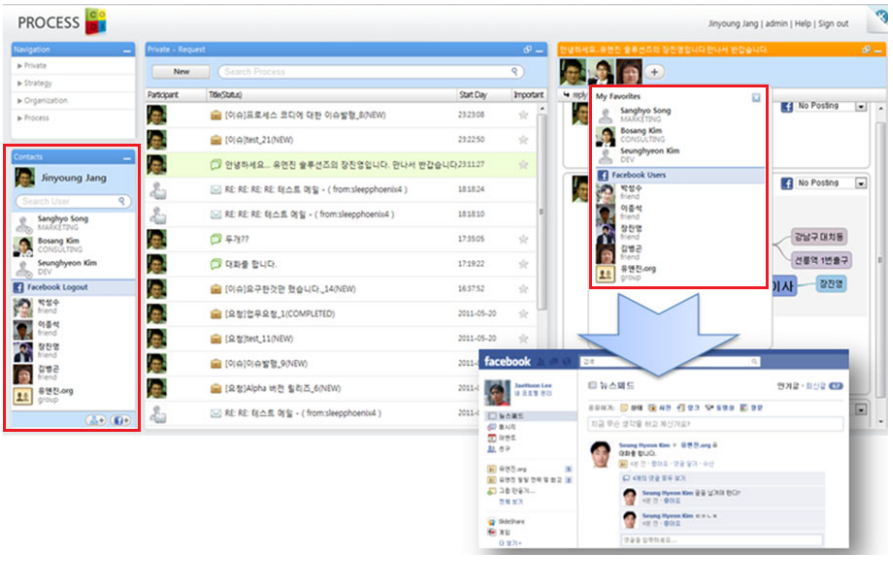


Fig. 2. Overview of ProcessCodi: it consists of four parts including navigator, contact list, worklist, and messenger

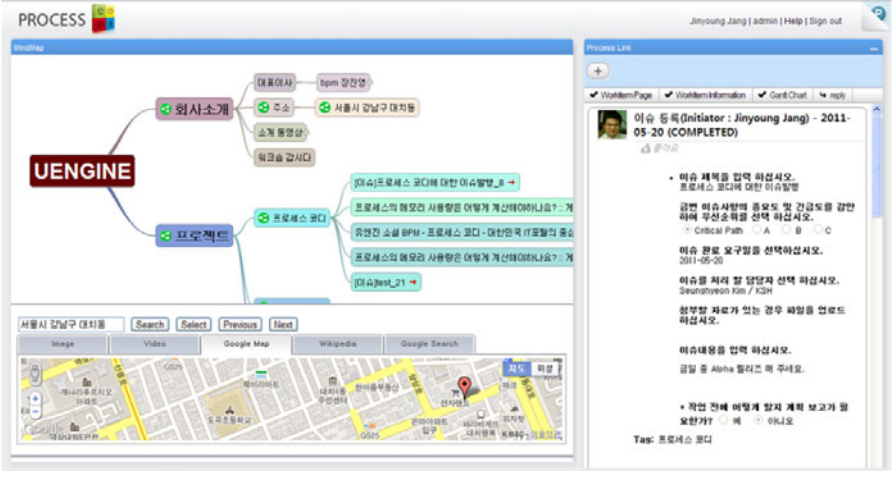


Fig. 3. Mind-map integrated in ProcessCodi as a knowledge management tool

3.3 Integration with BPMS

Business tasks are often to start from a time point when he/she gets an email. ProcessCodi provides an integrated work list including workitems, emails, chatting, and SNS communication. A business process can be triggered at the point of taking a workitem. Participants for the process can be added dynamically, and the whole contents are stored and managed as process instances.

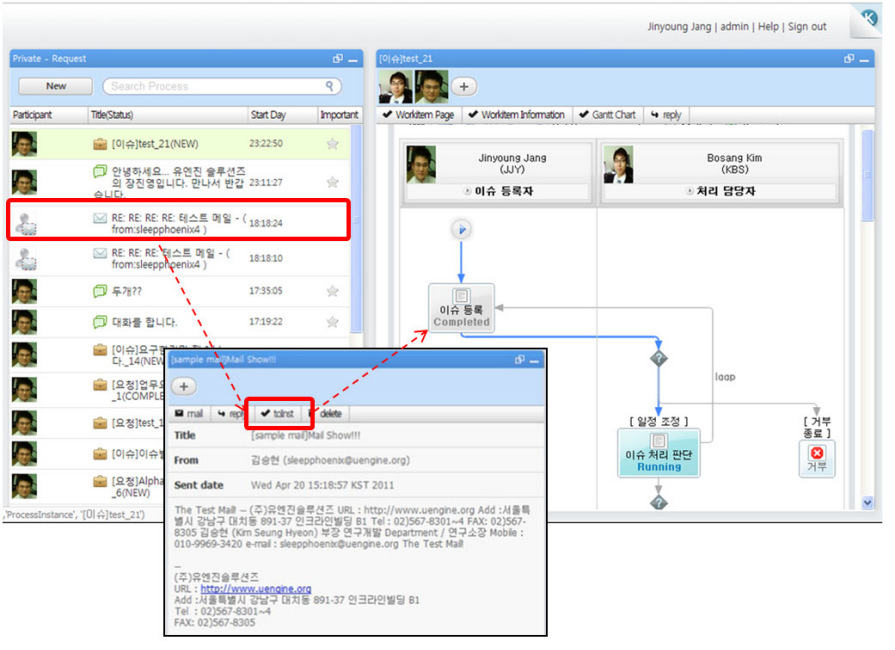


Fig. 3. Triggering of a business process from email, chatting, and SNS messages

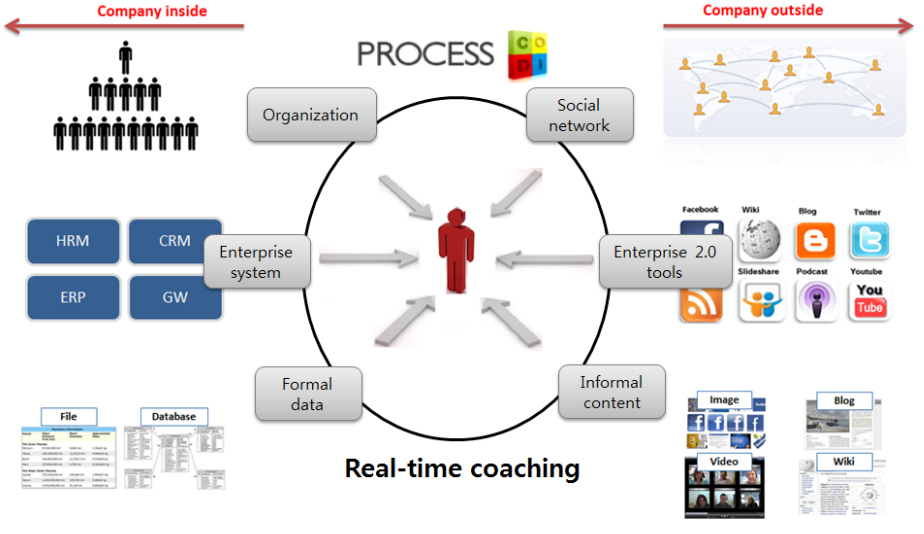


Fig. 4. The role of social BPM as a real-time coaching tool

3.4 The Role of Social BPM as a Smart Work Supporting Tool

Traditionally management information systems are limited in use of company inside due to the security problems. A social BPM is to satisfy the need of utilizing the resources of social networks and informal data as a knowledge base. One way of using social BPM is depicted in Fig. 4. In perspective of human networks, a user in ProcessCodi can involve not only employees in a company bus also experts or associates outside the company. In terms of tools, both the enterprise applications and social tools can be used in an integrated way. Likewise informal and formal data can be used and preserved in ProcessCodi.

The goal of ProcessCodi is to make workers more knowledgeable. We expect an intelligent knowledge tool such as real time coaching may be added to support right resources at right time. The real time coaching functions may be implemented by data mining or context awareness technologies. Ultimately, ProcessCodi aims to maximize the productivity of knowledge based smart workers.

4 Conclusion

In this study, we developed a Social BPM based on the concept of social business process patterns. Social business process patterns are a way to smart use of human resources and information in both the inside and outside of an organization. Our implementation: ProcessCodi is an integrated approach and tool, which provides a smart worker right resources at right time so that he/she can concentrate his/her own value added jobs. Consequently, ProcessCodi aims at integration of technologies and methodologies, and finally a total platform for smart works.

References

1. Llewellyn, N., Armistead, C.: Business process management: Exploring social capital within processes. International Journal of Service Industry Management 11(3), 225–243 (2000)
2. Li, X., Vrieze, P., Xu, L.: When Social Software Meets Business Process Management. In: 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology, Seoul, Korea (November 2009)
3. Koschmider, A., Song, M., Reijers, H.A.: Social software for business process modeling. Journal of Information Technology 25(3(15)), 308–322 (2010)
4. Hassan, N.R.: Using Social Network Analysis to Measure IT-Enabled Business Process Performance. Information Systems Management 26, 61–76 (2009)
5. Khoshafian, S.: MyBPM: Social Networking for Business Process Management. In: 2008 BPM and Workflow Handbook. Future Strategic Inc. (2008)

Integrated Process Assessment Framework to Be Enforced Functional Safety

Sun-Myung Hwang

Computer Engineering Dept.
Daejeon University, Republic of Korea
sunhwang@dju.kr

Abstract. This paper presented an integrated assessment framework including both process and functional safety. The integrated framework was designed by Automotive SPICE and ISO 26262-6 which are related on automotive software. Automotive SPICE, based on ISO/IEC 15504(SPICE), is a standard for assessment of automotive software and ISO 26262, based on IEC 61508 is an international standard for functional safety of road vehicle embedded software. Integration is proceeded in practice dimension and integrated base practices are modified or added, and included method and measure via safety integrity level, ASIL.

Keywords: Automotive SPICE, SW Safety, ISO/IEC 15504, IEC 61508, ISO 26262, Functional safety.

1 Introduction

Since CMMI(Capability Mutuality Model Integration) and ISO/IEC 15504: Software Process Improvement and Capability dEtermination (SPICE)appeared in 1990, process assessment of various domains has required such as embedded, automotive, aerospace etc., based on these models. Specially, in automobile domain Automotive SPICE has been made and used by Automotive SPICE SIG [1] and the one of most important characteristics of automobile domain is safety. A fault of it's imbedded a software is vary closed related to people's lives, the hardware as well as the software.

However, the case of Automotive SPICE assessment doesn't included to safety items.

So, in this paper, we propose Automotive SPICE with safety items of ISO/IEC 26262 can guarantee the safety of the software process assessment model of automobile area.

2 Related Research

Automotive SPICE [3] was created based on the ISO / IEC 15504 for automotive software process assessment model. It used in the automotive industry in order to review a variation of the characteristics used to fit the SPICE standard, and it has 31

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 384–390, 2011.

processes of SPICE. They are shown in Fig.1. Automotive SPICE has 6 capability levels from incomplete to optimizing and each process is assessed by process attributes.

IEC 61508 is standard about Functional safety of electrical, electronic & programmable electronic safety-related systems [4]. It is the standard for functional safety in electric, electronic and programmable electronic safety-related system. IEC 61508 defines 4 safety integrity levels, called ASIL.

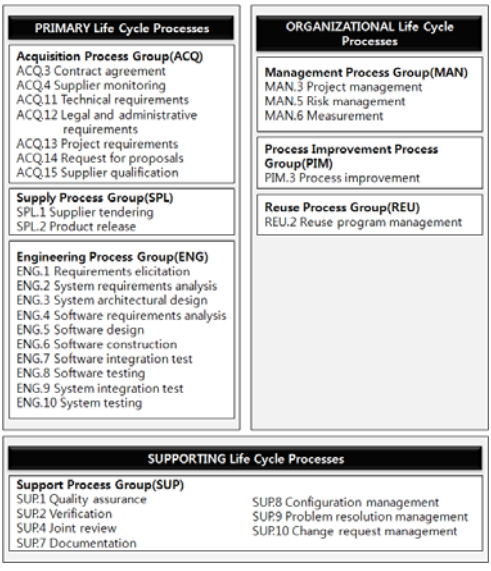


Fig. 1. Automotive SPICE Process

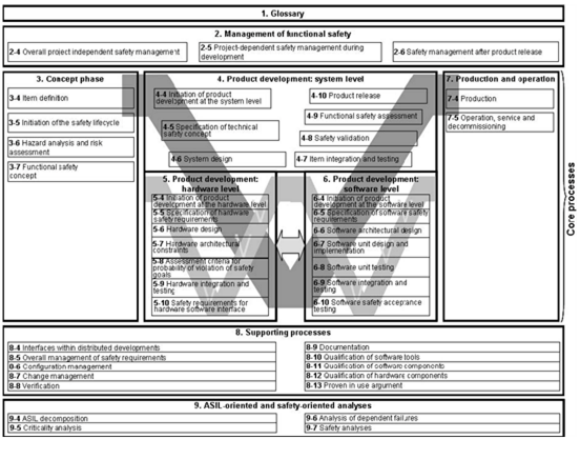


Fig. 2. The overall structure of ISO 26262

In the automobile, ISO 26262 is processing as Road vehicle - Functional safety [6]. The overall structure of ISO 26262 is shown in Fig.2. The level of functional safety is described to ASIL (Automotive Safety Integrity Level) There are 4 levels(A~D).

3 Integrated Framework Approach

Building strategy for establishment of the integration process Automotive SPICE and ISO 26262-Part 6 was planned as follows.: First, Integration of Automotive SPICE and ISO 26262's ability level of ASIL is concerned. In order to integrate two models, we propose 3 dimensional structure(x-axis:process reference model, y-axis: capability level, z-axis: ASIL) in Fig.3.

To include functional safety measures in Automotive SPICE, the following practices were added.

- Expansion of the existing base practice
- Additional safety practice in the existing base practice
- Extension of general Practices to achieve each capability level.
- Method indicates recommend degree according to ISO 26262 .

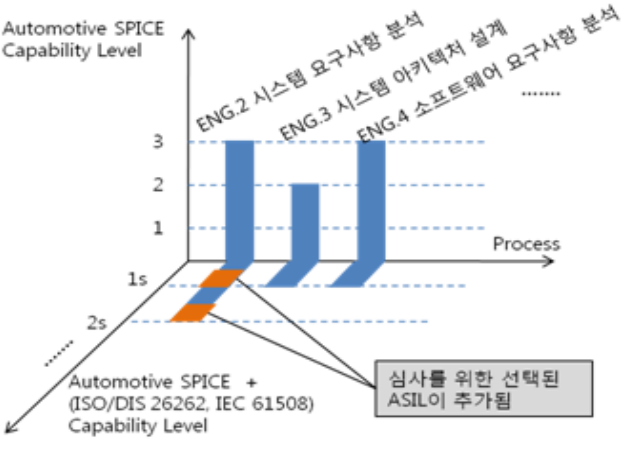


Fig. 1. Integrated assessment framework

To reflect these changes (Table 1) Automotive SPICE PAM 2.5, ENG.4 software requirements analysis process and the software safety requirements specification of ISO 26262-6 are mapped in Table 3.

Automotive SPICE ENG.4 results mapped the existing six of the eight-base practice changes are made due to the ISO 26262-6 and the new one was added to base practice.

Table 1. ENG.4 of Automotive SPICE

ENG.4 Software Requirements analysis	
BP	BP1: Software Requirements Identification BP2: Software Requirements Analysis BP3:The influence on the operating environment BP4: Software requirements, selection and classification of the priority BP5: evaluate and update the software requirements BP6: software requirements from system requirements to ensure consistency and bi-directional traceability. BP7: system architecture design and software from the bi-directional traceability of requirements to ensure consistency BP8: Software Requirements radio
WP	Software Requirements Specification Requirements trace rable

Table 2. Software safety requirements specification in ISO 26263-6

5. Specification of Software safety requirements	
practice	5.4.1 General 5.4.2 SW requirements for safety management in accordance with the overall safety requirements must be specified. 5.4.3 SW safety requirements allocated to the SW and SW technical and safety requirements must be derived from the safety requirements should be marked. ... 5.4.4 SW due to changes in the way to ensure the integrity must be specified. 5.4.5 SW technical safety requirements, safety requirements and system design requirements and should be consistent. 5.4.6 SW assigned in conformity with the technical safety requirements or restrictions can not be achieved, the change request should be made. 5.4.7 SW safety requirements between the HW and SW shall identify the dependencies for each SW. 5.4.8 Other than that specified in 5.4.2 features, if performed by an embedded SW, these functions explicitly, or a reference to the specification should be created. 5.4.9 SW safety requirements to enable the following: It should contain enough information. 5.4.10 SW safety requirements, if the function fails to violate safety requirements and ASIL should include all the features. 5.4.11 ... ... listed in Table 2 and Table 3, using the ways and means should be verified. BP5: evaluate and update the software requirements BP6: software requirements from system requirements to ensure consistency and bi-directional traceability. BP7: system architecture design and software from the bi-directional traceability of requirements to ensure consistency BP8: Software Requirements radio
WP	SW safety requirements specification SW Safety Requirements Verification Report

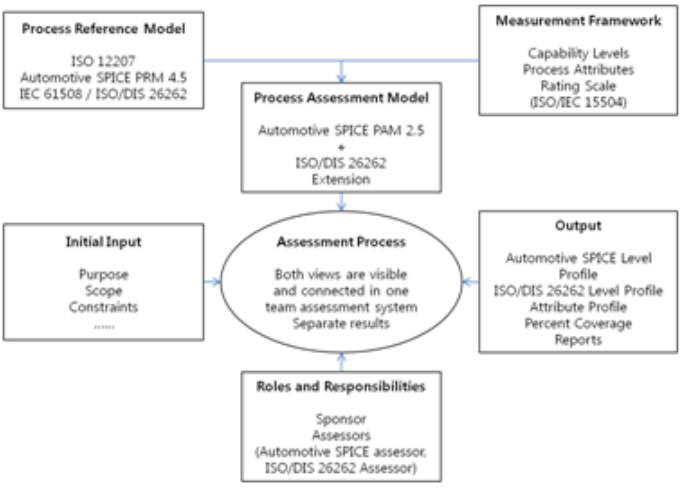


Fig. 1. Integrated framework for process

The integration of ISO 26262 and Automotive SPICE process assessment model framework (see Fig.4) can be expressed as [9].

Rating of each process in the same way in the SPICE Not achieved, Partially achieved, Largely achieved, Fully achieved evaluates each general practice.

Table 3. Mapping table

Automotive SPICE	ISO 26262-6	Integrated Base Practice
BP1	5.4.2 5.4.8	BP1: Software Requirements Identification Identification of the functional and nonfunctional requirements for the software as the basis for system requirements and system architecture, design, use, and software requirements are documented in the software requirements specification. Software safety requirements specification in accordance with all safety requirements are managed.
BP2	5.4.3 5.4.9	BP2: Software Requirements Analysis Technical feasibility, risk, and about the possibility of testing and analyzing software requirements are identified. Software safety requirements shall include the following considerations. a) system configuration and architecture b) HW and safety requirements, HW-SW interfaces and HW architecture c) the timing of the appropriate clearance d) External Interface e) SW and systems that affect the vehicle's operating mode

Table 3. (*continued*)

BP3	5.4.4	BP3: the influence on the operating environment Software requirements, system requirements, and the interface between the operating environment and other components, to define its impact on the safety integrity. Software safety requirements related to each of the software between the hardware and software shall identify the dependencies.
BP4	-	BP4: Software requirements prioritizing and classifying Identified and analyzed, prioritize software requirements for future releases, it maps to type.
BP5	5.4.6	BP5: evaluate and update the software requirements Cost, schedule and technical impacts on the impact of proposed changes to the software requirements and system requirements and architecture design to evaluate the changes. If the software is assigned to the technical compliance with safety requirements or restrictions can not be achieved, the change request should be made. Software requirements and software requirements for approval to renew the statement.
-	5.4.10 5.4.11	BP6: Software Requirements Validation If the function fails, the software requirements, safety requirements and a violation of ASIL sure you include all the features that will be verified. Verification of software requirements listed in Table 2 and Table 3, using the ways and means to verify.
BP6	5.4.5	BP7: The consistency and bi-directional traceability of software requirements from system requirements The software requirements including verification criteria and verification of system requirements that include criteria to ensure consistency. Consistent criteria for the verification and validation criteria that include the system requirements, including bi-directional traceability between software requirements are supported by establishing and maintaining.
BP7	5.4.5	BP8: System Design Software from ahkatekcheo bidirectional traceability of requirements to ensure consistency The software requirements including verification criteria and verification standards, including system architecture design to ensure the consistency of. Consistent criteria for the verification and validation criteria that include system architecture design, including bi-directional traceability between software requirements are supported by establishing and maintaining.

Table 3. (*continued*)

BP8	-	BP8: Software Requirements radio All interested parties needs and requirements for software updates to propagate to establish a communication system.
Integrated output		Software requirements specification (including software safety requirements specification), software requirements verification report ,Req. trace table

4 Conclusion and Future Works

In this paper, we proposed a integrated process assessment framework to bee enforced functional safety based on Automotive SPICE. The functional safety characteristics is referenced by ISO 26262 framework that reflects the level of the ASIL. By using this framework we can obtained the results created the practice of software for safety-related information in Automotive SPICE.

As future research, incorporation of Automotive SPICE and ISO 26262 is actually to be utilized for screening, as well as integration of general practice is needed.

Acknowledgement. This work was supported by the Security Engineering Research Center granted by Korea Ministry of Knowledge Ecomony.

References

1. “Automotive SPICE", http://www.automotivespice.com "
2. Panaroni, P.: "Safety in Automotive Software: an Overview of Current Practices" (2008)
3. Hoermann, K.: Automotive SPICE in Practice, pp. 1–10 (2008)
4. KS C IEC TR 61508:2009 “Functional safety of electrical, electronic & programmable electronic safety-related systems
5. Functional Safety and IEC61508“, http://www.iec.ch/functionalsafety "
6. ISO 26262-6:2009, "Road vehicles - "Functional safety - Part 6: Product development: software level"
7. Messnarz, R.: Integrated Automotive SPICE and Safety Assessments, 279–288 (2009)
8. Automotive SPICE SIG, "Automotive SPICE® Process Assessment Model" RELEASE v2.5 (May 2010)
9. ISO/IEC 15504:2004, "Information technology - Process assessment - Part 2:Performing an assessment"

Guideline for Moodle Customization

Seon Kyoon Park[1], Jung Suk Choi[2], and Jeong Ah Kim[3]

[1] Management Information Depart., Kwandong University,522 NaeKok Dong, KangNung, Korea
[2] Educational Technology Depart., Kwandong University, KangNung, Korea
[3] Computer Education Depart., Kwandong University, KangNung, Korea
refmeju@nate.com, east1910@nate.com, clara@kd.ac.kr

Abstract. The Moodle Customization requires correction of interface and functions to fit the characteristics of organization, operation method and characteristics of the students. There are User Management, Lecture Management, Functional Organization management and UX Management as the Moodle Customization methods. You can apply Moodle according to the characteristics of organization by customizing Moodle.

Keywords: LMS, customization, open source, product line.

1 Introduction

Recently, the open-source software are actively being used in various fields. While the software users usually think of Linux from OS class for the open-source software, these are applied by extending in various ways such as LMS (Learning Management System, hereinafter LMS), CRM(Customer Relation Management) or BPM(Business Process Management) as applications used for accomplishing a specific general purpose. Due to the extended spreading of e-learning and appearance of u-learning, the use of LMS is getting extended. However, it requires a lot of cost to adopt because a commercial LMS developed to fit each school environment must be adopted to build up the LMS. An open-source software LMS called Moodle can be utilized as the alternative.

Moodle is an abbreviation of Modular Object Oriented Dynamic Learning Environment which is one of LMS' making a modular object oriented and dynamic learning environment. In case of using Moodle, there are advantages of saving on costs because it is an open-source software and making corrections possible as desired because the code is open to public. There are also various learning resources even inside Moodle and more functions are provided using Plug-in or Filter. Moodle provides a stable support as it continues to get updated quickly. However, there is great shortage of the documents to be referred on the method of using Moodle while the prepared outcomes of development are unsatisfactory because it is an open source which is inconvenient to use and from the Korean culture of many developers participating in the development. There is also a great deal of difficulty because a user

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 391–396, 2011.

has to make corrections after understanding through code analysis as there is a lack of documents to be referred while making corrections.

In this paper, the requirements of Moodle customization with highest priority will be identified and the customization guidelines by each requirements will be presented to be used together with the school affairs information systems of each school in using Moodle which is an open-source software.

2 LMS (Learning Management System)

2.1 Trends

LMS is a system which manages the general details necessary for school affairs management such as checking attendance as well as managing the learning history management of tracking the grades and learning process of the students through an online network. LMS is a system for an integrated operation and management on the entire process of teaching-learning from the development and evaluation to the management of the contents that are necessary for learning.

As e-learning has become active, the importance of LMS is getting highlighted to have extended features provided and wide range of usage. The LMS of early stage was at the level of supporting simple blackboard contents, VOD contents or HTML contents. In order to gain a higher learning effect by accommodating the changing learning environment and various learning details, it is necessary to consolidate the learning management and support functions of LMS [1]. That is because a customized learning environment for the individual learning of students can be organized effectively as the function of LMS becomes improved.

The LMS currently developed in our country includes the record and learning management of students along with a function to transmit the learning details while also providing the function of connecting with other learning tools such as competence and learning skill management function, a function to analyzed the difference of learning skills, learning object management function and counseling system, LCMS (Learning Content Management System), academic ability diagnosis system or community system.

Recently, it is getting evolved as a form of solution that can actively substitute according to service type by making the learning management and other functions as solutions to be developed as a form of consolidating the relationship with, KMS(Knowledge Management System), groupware or collaborative solutions [2].

As commercial LMS' abroad, there are Blackboard and WebCT. The supply and spread of open-source software LMS' made up of voluntary participants by development is in progress.

2.2 Open-Source Software LMS

Moodle and Sakai could be called typical open-source software. Moodle can be easily installed and maintained while being able to organize the learning activity factors focusing on joint learning, collaborative learning and learning activity of students

based on constructivism in the educational perspective. Moodle can be operated by integrating or connecting with external system such as school affairs system, Wiki system, content management system, Social Networking System, e-portfolio, chat server, media server or web conference.

Sakai is an off source project started for the joint development of online learning system which had been used after being developed independently by University of Indiana, University of Michigan, MIT or Stanford University. Sakai is made based on Java. The instructor manages the learning materials such as assignments, chats or resources by making a website and .able to use by setting up the functions that are necessary in the lecture.

Moodle is able to access the code more easily than Sakai using a programming language called PHP and also enables easy maintenance. Moodle also supports more types of database to be used more easily.

2.3 Problem of Moodle Customization

In order to adopt LMS, the solution is generally changed on the interface and functions to fit the characteristics of organization, operation method and characteristics of students to apply which is not using the solution already developed as it is. Although Moodle is an open-source in which many developers have participated, the outcomes of development is unsatisfactory due to the nature of open source. As a disadvantage of having unsatisfactory outcome of development, there is a great deal of difficulty because the algorithm must be understood through source code analysis while customizing Moodle and make corrections as desired by the organization.

3 Moodle Customization Requirements

3.1 User Management

In order to use Moodle, the user must be registered first. For the user roles, there are administrator, lecture creator, instructor, assistant instructor and student. There are many methods of registering a user to Moodle including e-mail based authentication method and the method of administrator registering manually.

While adding a user to Moodle, having the instructors and students join or registering instructors and students one by one after the administrator looks at the academic affairs information would be waste of time. The information on instructors and students is already registered to the school's academic affairs system. It is necessary to bulk register the users to Moodle by importing only the necessary required information on these instructors and students.

3.2 Lecture Management

There are lecture setup, instructor registration and student registration as typical functions of lecture management. The method of setting up the lecture in Moodle gets to create a lecture inside a category after creating an administrator category. The

method of registering instructors or students to the lecture must also have the roles of each instructor and student assigned to the users by the administrator.

The administrator registering instructors and students for each lecture after creating lecture is a hassle because it requires a lot of time. The data related to lecture must be bulk registered to Moodle by importing only the necessary required information on the lecture, instructor in charge and students who have applied for lecture that are already registered to the student affairs system.

3.3 Functional Organization Management

In Moodle, there are learning resources and learning activities as various functions to support learning. There are screen, data, path, address, cover and IMS content pack for learning resources while there are submit assignment, chat room, database, brief survey, Wiki, research, complete learning, forum, quiz problems, dictionary, Scorm./ AICC(Aviation Industry CBT Committee) and cooperative learning, etc. You can also add new learning activities.

While using Moodle by applying in each school, using all of various learning resources and learning activities is a difficult thing. The functions must be managed to use only the functions necessary for each school among various learning resources and learning activities.

3.4 User Interface

Moodle has UI of unsophisticated design. Although the UI of Moodle can be used as it is, the users encountering this for the first time go through confusion from not being familiar with UI.

It is necessary to change the UI of Moodle for the convenient usage by users including the school logo, image, icon, font size, design and screen location of menu, etc.

4 Moodle Customization Method

4.1 Customization Techniques for User Management

The process of transferring users from the existing school affairs system for the convenience of user registration is shown as <Table 1>. The registration of user from Moodle database is user table while the path management on Moodle system and user is context table.

4.2 Customization Technique for Lecture

The process of transferring data from school affairs system for the creation of lecture, creation of category or convenience of instructor and student registration is shown as <Table 2>.

Table 1. User Migration Process

No.	Step	Detailed Activity	Reference
1	Change of user table attributes	. Change the default attribute values of user table	
2	Extracting data from school affairs system	. Import data such as ID, Password, Name and E-mail address, etc that are necessary categories as user information.	
3	Convert user information as data for input to the Moodle user table	. Enter the name Imported from the school affairs system by dividing into last name and first name	
4	Enter user data to Moodle	. Enter ID, Password, Name and E-mail address, etc that are minimum categories required for user registration . Enter context label, instant ID, path and depth that are necessary for path management of user	

Table 2. User Transfer Process

No.	Step	Detailed Activity	Reference
1	Change of lecture related table attributes	. Change the default attribute values of lecture related table	
2	Extracting data from school affairs system	. Import department name which is an item necessary as category information. . Import lecture name and lecture abbreviation that are items necessary as lecture information . Import ID and Name that are information of the instructor in charge of lecture or information of student taking the lecture.	
3	Enter data to the Moodle lecture related table	. Enter category name, upper category ID, category depth and category path that are minimum categories required for category registration. . Enter category ID, full name of lecture and lecture abbreviation that are minimum items required for lecture registration . Enter data on blocks and modules sued for lecture . Enter context data necessary for path management of category, lecture, block and module . Enter role ID, lecture ID and user ID, etc that are minimum categories required for user role registration	

4.3 Customization Technique for Capability Features

The process of managing function to use only the functions necessary at each school among various learning resources and learning activities in Moodle is shown as below.

The table managing the learning resources and learning activities in Moodle database is the modules table.

4.4 Customization Technique for UX

The process of changing Moodle for convenient usage by users is as follows.

For the change of UI in Moodle, execute the change of theme. To change the theme, select Module - Theme - Theme Settings page to select and change the theme.

To prepare and change UI properly for each school, the theme can be installed by preparing in the theme folder at the root where Moodle is installed. The following (Figure 7) is the file that must be drafted while preparing the theme.

5 Conclusion

In this thesis, the customization method in case of using Moodle which is an open-source software LMS in schools has been presented. The requirements on customization are user management, lecture management, functional organization management and UX management.

The user management and lecture management have been customized by controlling the database directly. In case of customizing by controlling the database directly, the data must be entered considering the relationship between each table. If the relationship between tables become deviated by not entering the data properly, there are cases where normal operation of Moodle is impossible. The functional organization management can be easily changed through administrator homepage. The UX management must apply design by creating the theme.

The improvement and research toward the method of customizing users and lectures using the functions of Moodle without controlling the database directly are the parts that must be performed. Also in case of UX management, the research toward the method of customizing for the convenient usage by users escaping from the existing frame of Moodle is also the part that must be performed. Once the research on user and lecture transfer or change of UX using functions becomes in active progress, we would be able to customize Moodle a little more easily.

Acknowledgments. This research was supported by National IT Industry Promotion Agency (NIPA) under the program of Software Engineering Technologies Development and Experts Education.

References

1. Na, H.: A Study on Model of Learner Oriented u-LMS. Ph.D thesis of SoongSil University (2008)
2. Nam, Y.: Implementation of LMS Interconnection Model for Inter-University e-learning Management. Ph.D Thesis of KangWon University (2009)
3. Moodle, `http://moodle.org`
4. Moodle user's community in Korea, `http://www.moodle.or.kr`

The Development of an Interactive Digital Textbook in Middle School English

Jeong-Im Choi[1], Heeok Heo[2], Kyu Yon Lim[3], and Il-Hyeon Jo[4]

[1] Department of Educational Technology, Kwandong University, 522 Naekok-dong, Kangnung, Kangwon-Do, South Korea
[2] Department of Computer Education, Sunchon National University, 315 Maegok-dong, Sunchon, Chonnam, South Korea
[3] Department of General Education, Ajou University, San 5 Woncheon-dong, Suwon, South Korea
[4] Department of Educational Technology, Ewah Woman's University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul, South Korea
choij@kd.ac.kr, hoheo@sunchon.ac.kr, klim@ajou.ac.kr, ijo@ewha.ac.kr

Abstract. The purpose of this study was to design and develop an interactive digital textbook in middle school English for both the classroom and individualized (differentiated) learning. Most of digital textbooks developed so far have been mainly focused on the digitization of text like an e-book, but not sufficiently utilized the interactive features of computer. Therefore, in this study, we attempted to find a way to develop the digital textbook to support students' individualized learning as well as classroom learning by accommodating various needs of individual learners and English education. For this, we devised the design principles for the development of digital textbook based on the theoretical review, which consisted of two dimensions of activities (Representation of information and Expression of understanding) and three categories of cognitive aspects (Recognition, Strategy, and Affection). Then, we applied them to the development of the English digital textbook. In this paper, we introduced the examples of the design principles applied in it and some implications for the development of digital textbook.

Keywords: digital textbook, development of digital textbook, instructional design, differentiated learning, UDL, middle school English.

1 Introduction

The advancement of information technology offers opportunities for new ways of creating teaching and learning resources including digital textbooks. Digital textbooks are defined as the digitalized forms of printed textbooks, which can be read, seen and listened through wired or wireless networks [4]. The digital textbook has maximized convenience and learning effectiveness with additional functions such as navigation, multimedia and learning supports with the advantages of the printed textbooks [2].

The digital textbook can be more dynamic and interactive in its transition from the traditional printed books to digital textbooks. It can be used to promote students'

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 397–405, 2011.

learning by integrating various types of multimedia resources for classroom lessons. Moreover, by quickly adopting the changes in knowledge and information, it can reduce the cost of publication as well as the burden of students' heavy backpacks. With that expectation, the Korean government has been developing digital textbooks since 2007. The Ministry of Education, Science and Technology has launched 'a digital textbook project in 2007 and run a pilot test at almost 100 schools in 2011.

As one of the projects, we have developed a middle school English digital textbook. This digital textbook, the first middle school digital textbook, was designed for the differentiated learning as well as regular classroom teaching/learning. According to the current middle school curriculum of Korea, schools are recommended to provide differentiated learning in English because English is regarded to be the most difficult subject in Korea. With the globalization and the growing importance of learning English however, students' English proficiency does not seem to improve [5].

We assumed that the digital textbook can be a solution to problems of English because digital textbooks can provide various multimedia resources and materials as well as the support for learning according to learners' understanding levels. However, the digital textbooks developed so far are basically designed to duplicate the printed textbook itself. Thus, we tried to find a way to design digital textbook to be a useful tool to support effective English learning in both individualized (especially differentiated) learning and classroom learning.

In this paper, we will introduce the design strategies to develop the digital textbook for both the classroom and individualized (differentiated) learning with some examples. And then, we will suggest some implications for the development of digital textbook.

2 Design Principles for a Digital Textbook

The purpose of this study is to develop a middle school English digital textbook for the use of both classroom and differentiated learning. For this, we attempted to find a way to accommodate various individual learners' needs as well as the needs in English education. We appreciated the UDL (Universal Design for Learning) as a foundational theory for our work because it provides conceptual framework as to how to design support for diverse students.

Universal Design for Learning (UDL) is originally developed for helping handicapped students (physically and mentally) in the special education to learn learning standard curriculum that designed for typical students in schools [6]. The key concept of UDL is to acknowledge the differences of learners and provide various teaching and learning methods to reflect those differences.

Rose & Meyer (2002) developed the UDL principles based on the neuro-science research suggesting that our brain consisted with three parts of recognition, strategic, and affective networks. Recognition networks are specialized to sense and assign meaning to patterns we see, so they enable us to recognize patterns and understand information. Strategic networks are specialized to generate and subsume mental and

motor patterns, so they enable us to plan actions and systematically act on information. Affective networks are specialized to evaluate patterns and assign them emotional significance, so they enable us to engage with tasks and learning. These three networks are structurally and functionally distinguishable but closely connected and functioning together. The implication of this brain research is that we need to support learning with various ways to promote these networks' functions and the way to reflect various learners' needs is to provide learning materials and contents in various ways [1].

We assumed that the functions of three brain networks can be applied to general education as well as special education. Generally, students use those networks to understand and integrate new formation into their knowledge. However, UDL was not enough to draw design principles for digital textbooks because it did not suggest enough information about the instructional design for digital textbooks. Thus, we reviewed instructional design theories that may be related to the principles of UDL and developed principles for the design of digital textbooks [3].

For the design of digital textbooks, we assumed that these factors can be considered in two ways [3]. First, we need to consider learners' cognitive process for understanding when we present information in the digital textbook. That means we need to design information to facilitate their learning by promoting the three brain networks, such as recognition, strategy, and affection. Second, we need to provide opportunities for students to express their understanding, by which we could know if learning is occurred. Even for the expression, we can consider each aspect of brain networks.

Based on these reflections, we have drawn a design guideline to support learning for a digital textbook as the Table 1. It consists of two dimensions of activities (Representation of information and Expression of understanding) and three categories of cognitive aspects (Recognition, Strategy and Affection).

Table 1. Design principles for a digital textbook

	Recognition	Strategy	Affection
Representation of information **Definition** The way to present and organize information	**Definition** The way to help the acknowledgement, recognition and understanding of information when the information is provided. **Strategies to support** R.R.1. Provide learning objectives. R.R.2. Represent content by using various media.	**Definition** The way to help the meta-cognition, strategic learning and reflection of information when the information is provided. **Strategies to support** R.S.1. Provide opportunities to grasp critical concept.	**Definition** The way to help the emotion, engagement and attitude on the information when it is provided. **Strategies to support** A.1.Provide multiple learning/teaching methods

Table 1. (*continued*)

	Recognition	Strategy	Affection
	R.R.3. Present the same content in different types of media. R.R.4. Emphasize important contents. R.R.5. Simplify complex tasks by using multiple steps. R.R.6. Provide prior knowledge and background information R.R.7. Provide authentic tasks and contexts. R.R.8. Provide various examples.	R.S. 2. Provide opportunities to summarize learning contents. R.S.3. Provide opportunities for practice. R.S.4. Provide opportunities for reflection. R.S.5. Provide various tools necessary for learning (note-taking, highlight, monitoring etc.) R.S.6. Provide opportunities to observe experts' performance.	A.2. Provide appropriate difficulties to learners' levels. A.3. Provide learner control for learning objectives, time, sequence, difficulties, etc. A.4. Provide user-friendly interface. A.5. Provide opportunities for success. A.6. Provide learning outcomes. A.7. Provide examples or tasks related to learners' experience. A.8. Provide opportunities for collaboration with peers.
Expression of understanding **Definition** The way for students to express their understanding	**Strategies to support** E.R.1. Provide learners with opportunities to demonstrate their understanding E.R.2. Provide various expression tools (writing, speaking, drawing, etc.) E.R.3. Provide various communication tools for interaction among participants. E.R.4. Provide opportunities for learners to see the result of response.	**Strategies to support** E.S.1. Provide prompting to promote learners' thinking. E.S.2. Provide opportunities for learners to share and compare their results E.S.3. Provide appropriate cognitive support for learners' levels.	

Note. First letters of each category are used for numbering. For example, R.R.1 represents the strategy under the category of Representation of information by Recognition.

3 Development of English Digital Textbook

The English Digital Textbook was developed in a Window OS environment with Flash applications and web-programming. It includes 12 lessons, each of which consists of 8 class-hours in middle school English. The Digital Textbook is delivered through a learning platform that provide learners involved in the use of the Digital Textbook with on-line services, learning and navigation tools to promote their successful performance. All of digital textbook developed by Korean government is serviced with this learning platform.

3.1 Listen & Speak

The 'Listen & Speak' section is developed for learners to practice listening and speaking in English. In listening practices, each activity requires learners to listen to the spoken version of the given sentences first and then to respond to questions related to it. In this section, we tried to provide various supports for learning. Fig. 1 shows an example of listening practice and scaffolds to facilitate students' learning.

Fig. 1. A screen shot of listening section in the Digital Textbook

The graphic icons(boxed) at the left-bottom of the screen include buttons for play to listen. It includes speed control buttons, hints, and script buttons as well as play button in order to allow students control the learning process according to their abilities. (*E.S.3. Provide appropriate cognitive support for learners' levels; A.3. Provide learner control for learning.)* While learners can see the script of a listening text in a pop-up window by clicking the 'Script' button, they can look at a Korean translation of the text in the script window by clicking the 'Kor' button. Students can see those scaffolds only when they had chosen incorrect answer. Also, students can read the question by Korean by clicking the button(boxed) in the right upper side of the screen.

Fig. 2 shows a screen shot of speaking section in the 'Listen & Speak' unit. In speaking section, learners can choose three options for speaking practice as shown in the right side of the screen (boxed). When click the 'Listen' button, learners can listen the conversation with animated graphics. (R.R.2. Present content by using various media; A.4. Provide user-friendly interface; R.S.6. Provide opportunities to observe experts' performance.) When clicking the 'Listen and Repeat' button, learners can choose a character of the scenario and participate in the conversation. If the learner has difficulties to understand the conversation, they can look at the script by clicking script button at the left-bottom of the screen. (E.S.3. Provide appropriate cognitive support for learners' levels.) The 'Talk with a foreigner' button shows a video of a native speaker. Learners are asked to speak with the foreigner to complete the conversation. (A.7. Provide examples or tasks related to learners' experience.)

Fig. 2. A screen shot of speaking section in the Digital Textbook

In speaking section, learners can also practice speaking with their friends at a distance using an on-line video chatting tool.(A.8. Provide opportunities for collaboration with peers.) In addition, video clips showing the shapes of a native speaker's mouths as they pronounce key vocabularies and expressions in each lesson were presented to help learners improve their pronunciation and speaking skills.(R.S.6. Provide opportunities to observe experts' performance.)

Learners can choose some learning tasks according to their English proficiency. The learning tasks are classified into low, middle and high levels, which is designated by Y (yellow), B (blue) and O (orange) in the upper-right hand corner of screen. The differentiated learning tasks allow learners to engage in learning with appropriate learning difficulties. (*A.2. Provide appropriate difficulties learners' levels.)*

3.2 Read & Do

The 'Read & Do' section is designed to develop reading skills in English. Reading requires lots of cognitive strategies to understand. Therefore, in this section several tools and buttons were designed to support reading strategies. First of all, before the presentation of main reading text, learners are asked to think about the theme by presenting prompting questions and pictures as a pre-reading stage. (*E.S.1. Provide prompting to promote learners' thinking; R.R.6. Provide prior knowledge and background information.*) After that, main reading text is presented with several scaffolding buttons for strategies on the right side (boxed) of the screen in Fig.4.

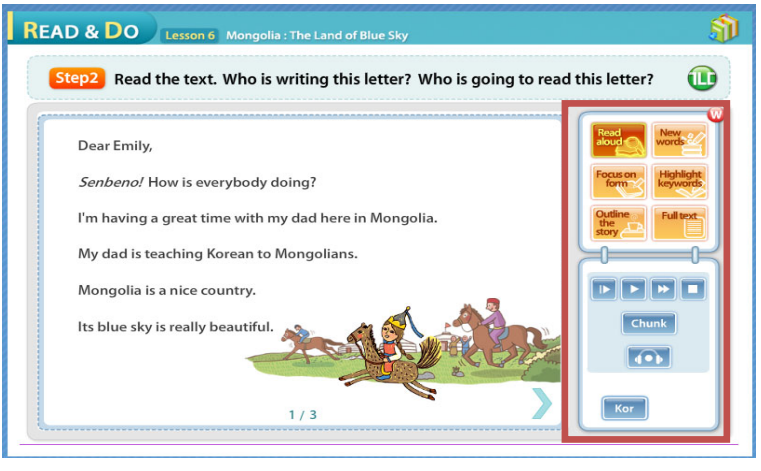


Fig. 3. A screen shot of 'Read & Do' section in the Digital Textbook

For example, when learners click the 'Read aloud' button, the paragraphs are spoken by a narrator. Learners can control the speaking pace using video buttons in the middle of the right side of the screen. Learners also can see the segment for effective reading by clicking 'Chunk' button under the video buttons. Learners can read by themselves by clicking 'headset' button. (*R.R.2. Present content by using various media; R.S.6. Provide opportunities to observe experts' performance; A.3. Provide learner control for learning*)

'New words' and 'Focus on form' buttons present new words and language forms by underlining in the reading text. When learners click the 'Highlight keywords' button, they can use a marker and an eraser to highlight meaningful sentences and expressions in the text. 'Outline of the story' button shows the whole structure of the contents. Even though the reading text is presented basically on separate pages, learners can look at the whole text on one page by pressing the 'Full text' button.These buttons were designed to help and support learners to comprehend reading text by stimulating strategy and affection. (*R.S.1. Provide opportunities to grasp critical concept; R.S.5. Provide various tools for necessary for learning.)*

Another additional tool for reading is the dictionary. The 'W'button on the right-upper corner of the screen shows the meaning of the word when learners click the words that they don't know after clicking the button. (*R.S.5. Provide various tools necessary for learning.*) Virtually, this button works not only in the reading section, but also in the whole section of the textbook. After reading, some questions related to the reading text are provided, which allows learners to express and check their understanding. (*E.S.1. Provide learners with opportunities to demonstrate their understanding; R.S.4. Provide opportunities for reflection.)*

3.3 Think & Write

Writing skills are enhanced in the 'Think & Write' section. Writing is a complex problem solving process, which requires planning, transcribing and reviewing steps [14]. Thus, in this section, we attempted to support learners' learning by presenting authentic tasks that they may use in real-life writing and provide sub-steps to complete the learning task. For example, learners can practice writing a letter, following several sub-steps pre-designed in the Digital textbook. After finishing all the sub-steps they can get the outcome of their learning which is a completed letter. (*R.R.5. Simplify complex tasks by using multiple steps; R.R.7. Provide authentic tasks and contexts; R.R.8. Provide various examples; A.6. Provide learning outcomes; A.7. Provide examples or tasks related to learners' experience.*)

3.4 Let's Check

'Let's Check' section is developed for assessing learning outcomes. Several test items with various levels of difficulty require learners to apply and integrate the four English skills: listening, speaking, reading and writing. No cognitive scaffolds are available while they are answering questions in this section. The results are shown after the learners have answered all the questions. Once all the questions have been solved, learners can get feedbacks and try to answer the questions again. (E.R.1.Provide learners with opportunities to demonstrate their understanding; E.R.4. Provide opportunities for learners to see the result of response.)

4 Discussion

In this paper, we have introduced the strategies to develop middle school English digital textbooks to support both classroom learning and differentiated learning. We attempted to find theoretical principles for the design of a digital textbook and to apply specific design strategies into the digital textbook. The design guideline was a useful tool to communicate as to how to design the contents and support available for the instructional designers, subject matter experts and programmers. It also served as a guide to contemplate the problems and difficulties that students confront when they study English in order to design the required support.

According to our pilot test, students had no difficulties to use the design features of the digital textbook in general. However, some features were not sufficiently utilized

as we intended due to the lack of technical support (such as the small size of the window, type-in errors) and lack of motivation towards the use of objectives and warm-up questions. This suggests that we need a further analysis on what features students choose to use in the respective situations. Moreover, since some features of this program were designed for the classroom use (such as learning objectives, warm-up questions, some strategies for reading, etc.), the usability test in the classroom teaching situation should be conducted.

References

1. Bremer, C.D., Clapper, A.T., Hitchcock, C., Hall, T., Kachgal, M.: Universal design: A strategy to support students' access to the general education curriculum. In: National Center on Secondary Education and Transition (NCSET), University of Minnesota (2002)
2. Byun, H., Choi, J., Song, J.: Research on the development of Electronic textbook prototype. Journal of Korean Educational Technology 22(4), 1–24 (2006)
3. Choi, J.I., Shin, N.: A digital textbook design principle based on the Universal Design for Learning. Journal of Korean Educational Technology 25(1), 29–59 (2008)
4. KERIS: Special issue: Digital textbook leading future education. pp. 1–15, 2007 ICT in Education White Paper, (2007)
   http://www.udeducation.org/teach/teaching_techniques/bowe.asp
5. Kim, J.W.: Problems in teaching English and effective learning methods. English Language & Literature Teaching 12(3), 167–186 (2006)
6. Rose, D.H., Meyer, A.: Teaching Every Student in the Digital Age: Universal Design for Learning, ASCD (2002),
   http://www.cast.org/teachingeverystudent/ideas/tes/

Construction of Online Behavior Monitoring System

SeHoon Kim[1] and SeungYoung Choi[2]

[1] Management Informaiton Depart., Kwandong University, 522 Naekok, Gangneung, Korea
[2] Computer Engineering Depart., Kwandong University, 522 Naekok, Gangneung, Korea
SeHoonKIM.s@gmail.com, boromi@gmail.com

Abstract. Recently, there is a trend of each school being equipped with class behavior analysis rooms for the improvement of class capability. However, many problems exist in the classroom monitoring system such as high cost, limitation of time and space on demonstrator and analyst, ineffective feedback method or difficulty of post self-analysis. In this paper, a method to reduce the cost of installing physical space and overcome the limitation of time for participation of specialists using class behavior analysis system design.

Keywords: Software product line, online behavior monitoring, evaluation.

1 Introduction

Recently, the universities are at a trend of setting up sample interviews rooms for developing employment capability or class behavior analysis rooms for developing classroom capability. The sample interview room is a place to analyze on the behavior during an interview. The structure is a room where an equipment to record video and sound as a method of coming out with a recorded video after the student goes in here alone to have a monologue. The class behavior analysis room is a place to analyze on the behavior during class. The structure is a room where an equipment to record video and sound with One-way mirror is installed as a method of performing analysis and demonstration at the same time after the class demonstrator, student and analyzing specialist enter in the room.

However, there are several limitations in currently operated sample interview rooms and class behavior analysis rooms. In case of sample interview rooms, there is a problem of analysis being ineffective due to a feedback by someone who is not a feedback specialist. In case of class behavior analysis room, there are disadvantages of requiring high cost for setup, many class behavior analysis specialists having to observe the class and the point of presenting a feedback on class behavior not being clear.

Accordingly, building up a real-time online system was determined as necessary in order to reduce the physical installation cost, have many analysis specialists participate in the evaluation as well as making the time of recorded video and presenting feedback same. In this paper, the goal is set as reducing physical installation cost by setting up a class behavior analysis room, raise the effect of class behavior analysis by increasing the possibility of participation by various analysis

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 406–412, 2011.

specialists and increasing the effectiveness of class behavior analysis by enabling reference on the specialist feedback along with demonstration video during post self-analysis.

2 Background

2.1 Importance of Class Behavior Analysis

Class behavior analysis is an activity of getting feedback from colleagues, students and specialists on the behavior factors of instructor for the purpose of improving the quality of class. If education is called a planned effort of trying to change the human behavior characteristics, the quality of education could be called the quality of class and the teacher is at an important position of determining the quality of class [1]. There are setup of education objectives, content organization, explanation, control & management of environment and control of student behavior, etc [1]. The class activity of a teacher who has received a feedback on class behavior analysis is greatly improved compared to the teacher who has not done so [2].

However, it is difficult to analyze and feedback on the teacher behavior factors through self analysis. Also in cases other than self-analysis, the colleague rating and counselor rating are more effective on the class behavior improvement than the student rating [2]. Therefore, many class behavior analysis specialists must present feedback by observing the teacher's class. But there is a limitation of having to adjust the schedule between teachers and class analysis specialists in having many class behavior analysis specialists participate in the class.

2.2 Class Behavior Analysis Method

2.2.1 Class Behavior Analysis Room

Until now, the class behavior analysis rooms have been set up at schools of each level based on specialized businesses. A classic class behavior analysis room is made up of class demonstration room where actual class demonstration is performed, class control room where class demonstration is recorded and class analysis room where class behavior analysis specialists observe class over One-way mirror.

The class analysis in a class behavior analysis room had disadvantages of having difficulty in identifying detailed class atmosphere taking place at a class demonstration room and impossibility in identifying the reaction of students by the class behavior analysis specialist observing in class analysis room. In a class behavior analysis room equipped with hi-tech equipment, many auto-tracing cameras and large monitors get installed. A class behavior analysis room equipped with hi-tech equipment was made to see the interaction between teacher and students more in detail than the classic class behavior analysis room while enabling self analysis in the future by saving the details of class demonstration. But the problem of having to make a class behavior analysis specialist participate and the problem of having high initial investment costs such as One-way mirror installation, etc still remain even in case of a class behavior analysis room equipped with hi-tech equipment.

2.2.2 Diversification of Analysis Method

Computer Assisted Self Supervision. Self supervision means the teacher planning, carrying out and evaluating the class by oneself with a purpose of improving the class. In such self supervision, there is a computer assisted self supervision using computer as assisting medium. The computer assisted self supervision is the method of actively supporting the self supervision of a teacher through overall support of human and physical environment based on computer[3].

For the computer assisted self supervision systems, there are 「KICE Teaching-Learning Development Center [4]」, 「Seoul Metropolitan City Education Office Class Support Team [5]」 and 「Edunet[6]」 site, etc [7]. These websites are the method of giving feedback once the recorded class video is registered online after the class demonstrator requests an online analysis.

While these websites have advantages of being able to perform even without a class behavior analysis room and not requiring a class behavior analysis specialist participate in the class, there are disappointments of having difficulty in identifying the precise class atmosphere because the class behavior analysis specialist did not participate in the actual class, unable to find out the precise feedback details because feedback is given online and not being able to find out the time of feedback by a class behavior analysis specialist.

Class Behavior Analysis Program. The class behavior analysis program supports observation and analysis of a teacher's class using computer [8].

The class behavior analysis program has AF (advanced flanders), student preference analysis, student task concentration analysis and analysis by student behavior factor as its main contents supporting a systematic and scientific analysis.

2.3 Directions

Although the class behavior analysis room is equipped with hi-tech equipment, having to invite many class analysis specialists during same time period and not being able to find out the time of class analysis specialist giving feedback while performing post self analysis still exist. In this paper, the following things will be improved.

First, the necessity of One-way mirror installation for physical classification between class demonstration room and class analysis room is eliminated. It enables physical classification between class demonstration room and class analysis room by transmitting the video filmed with many auto tracing cameras installed at the class demonstration.

Second, the necessity of on-the-spot participation by a class behavior analysis specialist is eliminated. The class behavior analysis specialist is able to perform class behavior analysis online even without participating in the class behavior analysis room where the actual class demonstration takes place.

Third, have the feedback of class behavior analysis specialist referred at the time of being pointed out during post self analysis. Making into a system, the reference of specialist's opinion for improvement is made possible by searching the point of feedback while performing post self analysis.

3 Research Method and Details

3.1 Domain Engineering

3.1.1 Definition of Requirements

The class behavior analysis room requirements are as follows.

First, the class demonstration video must be seen online realtime, Second, the class behavior analysis specialist must be able to analyze online even without participating on the actual spot. Third, the class behavior analysis specialist must be able to record the opinion of specialist on the class demonstration details instantly during the demonstration. The video and the video shown by broadcasting online must be synchronized. Fourth, the feedback data and class performance video must be saved after being matched. Fifth, the class demonstration video must be shown at the point of presenting feedback if the feedback is selected during post self analysis.

The usecase diagram of class behavior analysis room derived by the requirements defined above is shown as (Figure 1).

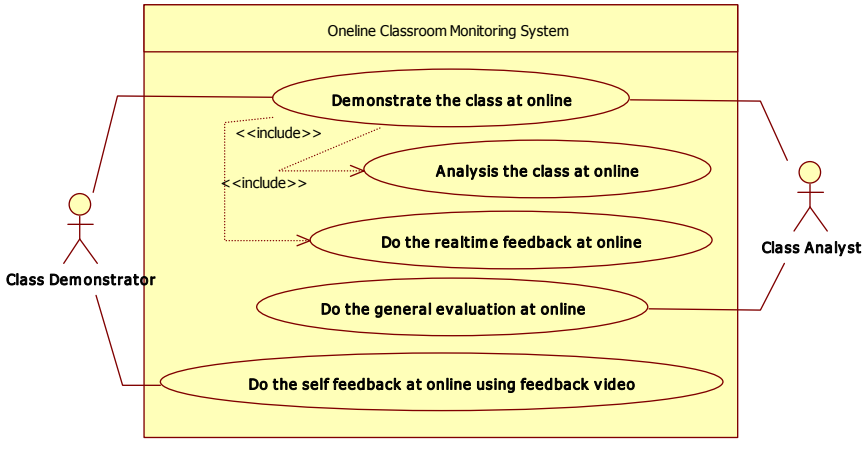


Fig. 1. Usecase Diagram

3.1.2 Architecture

For the class analyst to analyze the class demonstration video even without participating in the class demonstration, the class demonstration video must be shown online. Also, it was designed for the class analyst to prepare real-time feedback while watching the class demonstration video and perform general evaluation after the class demonstration is over.

There are relay server and analysis server. Relay server is a server in charge of role to transmit the video filmed from one camera real-time to the analysis server. Analysis server is a server in charge of role to provide online service after saving the video transmitted from the relay server. The reason for dividing server into two types is to place server role sharing and a possibility of extension on the number of

cameras. Through this, the system was designed to enable filming of class even at a place which is not a class behavior analysis room.

The video filmed from the class behavior analysis room transmits the class demonstration file real-time to the Media Server through Media Encoder. The Media Server saves the received flash file and online broadcasts the received flash file real-time.

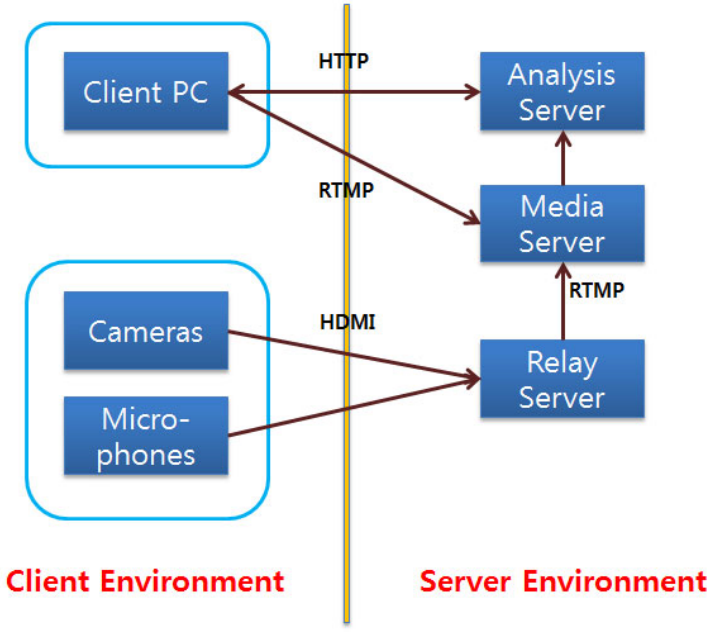


Fig. 2. Architecture

3.1.3 Implementation

The greatest problem in implementing system is transmitting video files with large volume. That is because the actual class demonstration screen and broadcasted class demonstration screen are not synchronized in case the volume of video file is large. So it is not transmitted as video file and transmitted by converting into flash file. In this system, the Flash Media Live Encoder was used as Media Encoder and used Wowza Media Server as Media Server. The class demonstration video is transmitted to the Wowza Media Server from the Flash Media Live Encoder. The class analyst gets to analyze and feedback on the screen shown below (Figure 3) by logging into the class behavior analysis system on the web.

In this system, the analysis can be performed with a maximum of 6 cameras. It was implemented to so that the currently viewed screen can be switched with existing screen by selecting a different camera screen in order to switch to a different viewpoint of camera while the class analyst is analyzing the class. Once the class demonstration ends, the class analyst gets to perform a general evaluation (Figure 3).

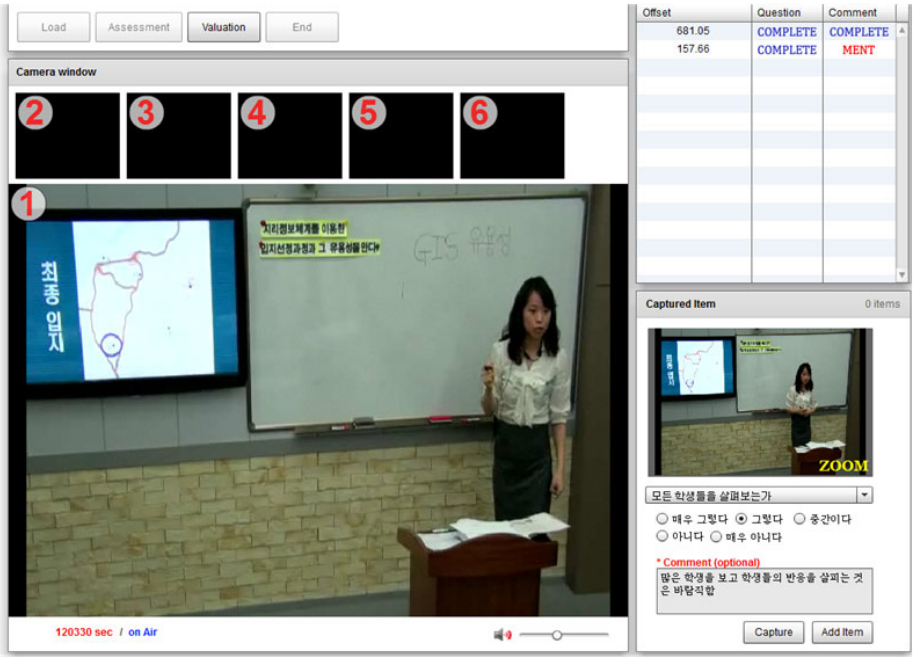


Fig. 3. Analyze Screen

4 Study Results

This study was started in order to diversify the method in which a class analyst can participate in the class. To make this possible, the class demonstration video was filmed to be shown online, enabled preparation of feedback realtime during class demonstration and enabled general evaluation after the class demonstration. Using this system, the class details can be analyzed by having class analyst participate with various methods and the class demonstrator is able to received effective feedback during self analysis. Furthermore, this would be helpful for front line schools or universities trying to set up a class behavior analysis room with little cost.

However, there are problems of a class analyst not being able to see the screen by Zoom-in and the problem in which the parts that class analyst can miss exists because only one large screen is shown and five remaining small screens are shown while performing analysis. These areas still need improvement. Also, it is necessary to apply the software product line engineering methodology to effectively cope with the diversity of requirements by each school. In other words, we are planning to research an architecture that has reorganized by school characteristics after analyzing the Feature that a cyber class behavior analysis room has based on the basic requirements and basic architecture design identified through this research.

References

1. Ryu, J.S.: A Study on the Component Factors of Teachers' Effective Teaching Behaviors. Ph.D thesis of Wonkwang University (2008)
2. Yoo, S.K.: The Effects of Different Feedback Styles on Teaching Behaviors of Teachers. Master Degree Thesis of Korea National University of Education (1994)
3. Kim, K.H.: The Effects of Compter Assisted Self-Supervision on the Improvement of Teaching Skill and Teaching Efficacy for Beginning Teachers. Ph.D thesis of Pusan National University (2003)
4. Teaching and Learning Center of Korea Institute for Curriculum and Evaluation, http://classroom.kice.re.kr
5. Seoul Metropolitan Office of Education, http://sooup.ssem.or.kr
6. Edunet, http://www.edunet4u.net
7. Lee, H.K.: Design and Development of a Teaching Behavior Diagnosis System for Self-Supervision. Master Degree Thesis of Korea National University of Education (2008)
8. Edusugar, http://www.edusugar.com

A Bootstrapping Method for Learning from Heterogeneous Data

Ngo Phuong Nhung and Tu Minh Phuong

Department of Computer Science,
Posts & Telecommunications Institute of Technology,
Hanoi, Vietnam
nhungnp@ptit.edu.vn, phuongtm@ptit.edu.vn

Abstract. In machine learning applications where multiple data sources present, it is desirable to effectively exploit the sources simultaneously to make better inferences. When each data source is presented as a graph, a common strategy is to combine the graphs, e.g. by taking the sum of their adjacency matrices, and then apply standard graph-based learning algorithms. In this paper, we take an alternative approach to this problem. Instead of performing the combination step, a graph-based learner is created on each graph and makes predictions independently. The method works in an iterative manner: labels predicted by some learners in each round are added to the labeled set and the models are retrained. By nature, the method is based on two popular semi-supervised learning approaches: bootstrapping and graph-based methods, to take their advantages. We evaluated the method on the gene function prediction problem with real biological datasets. Experiments show that our method significantly outperforms a standard graph-based algorithm and compares favorably with a state-of-the-art gene function prediction method.

Keywords: Semi-supervised learning, multiple data sources, graph-based learning, bootstrapping, multiview learning.

1 Introduction

Most existing machine learning algorithms are designed to work in the setting where data come from a single source. In many real-life applications, however, there may be multiple data sources related to the learning problem. In such cases, it is desirable to effectively explore and exploit the sources of information simultaneously to make better inferences. This kind of learning problem is often called multiview learning where each view corresponds to a data source [16].

There are many examples of multiview learning in different application domains. In protein classification problem, for instance, proteins are described by several types of biological data such as protein interaction networks and gene expression profiles. In the former description, the data can be naturally represented by a graph. In the latter description, each protein is represented as a vector of real numbers. Other types

T.-h. Kim et al. (Eds.): FGIT 2011, LNCS 7105, pp. 413–422, 2011.

of data such as co-localization of proteins in cells, structural similarities, and protein sequence data are also possible. Since different data sources describe different aspects of protein functions, utilizing all of them is beneficial for improving classification accuracy.

In this paper, we consider the setting in which each data source is represented as a undirected graph, and graphs are used together in learning and prediction [11]. For networked or linked data such as protein interaction networks or web page hyperlinks, it is straightforward to build graph representations. For vectorial data such as gene expressions or web text, one can build graphs in which vertices are objects (proteins, web pages) and weighted edges represent measures of association or similarity between the vector representations of the objects.

When data sources are represented as graphs, a possible strategy is to combine the information from multiple graphs before learning/inference. There are two main approaches for graph combination: summation of graphs [11,16] and summation of spectral kernels [2,12]. In this work, we take an alternative approach. Instead of combining graphs or kernels, each graph is used to train a classifier that makes its own predictions about the labels of (unlabeled) instances. The algorithm works in an iterative manner. In each learning round, labels predicted by some of classifiers are added to the training set of the remaining classifiers. Here we assume the semi-supervised learning setting, i.e. besides training instances with known labels we are also given unlabeled instances, for which predictions should be made.

The proposed method can be seen as a combination of two popular semi-supervised techniques. First, since the method uses labels predicted in previous iterations to augment the (labeled) training set in next iterations, it resembles the bootstrapping [1] or, more precisely, the co-training algorithm [5]. Second, in each iteration, the method uses a graph-based semi-supervised learner [15], thus it is graph-based. A similar work that tries to bring together bootstrapping and graph-based semi-supervised methods has been proposed in [13]. In this paper, we extend previous work for the case with multiple views and verify the usefulness of the proposed method in the problem of predicting gene function from heterogeneous biological data sources and compare with other graph-based methods.

Related Work. Semi-supervised learning has attracted a considerable research interest with a number of works published [6,17]. Bootstrapping and graph-based methods are among the main approaches to this type of learning problems.

In the simplest version of bootstrapping, also known as self-training, a classifier trained on labeled instances is used to make predictions for unlabeled instances [1]. A subset of the newly labeled instances is added to the previous training set and the model retrains. An important class of bootstrapping algorithms is co-training [5], in which two classifiers are trained on two different views of the same learning problem, and the predictions of either of classifiers on the unlabeled instances are used to augment the labeled set of the other. Co-training has been extended in several ways [3,5,17], mainly to relax some assumptions of the original algorithm, and to select good predicted labels to update the training set. In a recent work [14], Zhang and Zhou proposed a method to select confidently predicted labels. Our method is in essence an extension of their method for the case where there are more than two views.

Graph-based algorithms form another popular family of techniques for semi-supervised learning. In these methods, a graph spanning all labeled and unlabeled

instances is constructed. Then, different algorithms are applied to assign scores to nodes so that score differences between neighbor nodes are smooth, together with other constrains [15]. Graph-based algorithms have shown to deliver accurate predictions in a number of real-life problems [7,8,12]. Recently, there have been attempts to bring together co-training and graph-based methods into a unified framework [13], which appears to be promising. Here, we extend this line of research for the cases where more than two data sources present.

There are a number of works on learning from heterogeneous data sources, showing growing interest in this machine learning problem. A natural approach is to create a kernel or graph for each data source and then combine the kernels by taking the (weighted) sum of them [2,10,11,12,16]. Unlike these methods, we do combine kernels/graphs a dedicated step. Instead, we train a classifier on each graph and make predictions in the bootstrapping manner.

2 Bootstrapping with Multiple Graphs

Suppose we are given a set X of n examples, some of which have labels from $Y = \{-1, +1\}$, the others are unlabeled. By L and U we denote the set of labeled and unlabeled examples respectively so that $X = L \cup U$. For simplicity, we now assume that the examples are represented by m undirected graphs. We will describe how to construct graphs from data sources later. All m graphs share the same set of n vertices corresponding to n example. Graph number k is given by specifying its weight matrix W_k of size $n \times n$ where element w_{ij} is the weight of edge ij connecting i-th and j-th vertices. The goal is to predict the labels for examples from U given the graphs.

2.1 Algorithm Overview

The proposed algorithm is similar to the standard bootstrapping algorithm. Given m graphs on a set of labeled and unlabeled training examples, the algorithm proceeds in a number of rounds, in each round it performs the following steps:

1. Perform graph-based learning and make predictions for unlabeled examples on each individual graph.
2. Find confidently predicted labels by combining predictions from all m learners.
3. Add the confidently predicted labels in step 2 to the list of labeled examples and proceed to the next round.

In the following sub-sections we describe each step in detail.

2.2 Graph-Based Semi-Supervised Learning

Given a graph with weight matrix W, we use a semi-supervised learning technique described in [15] to get the label scores for unlabeled examples. The goal of this learning technique is to update score for each example iteratively. This technique is based on the assumption that nodes close to each other should have similar labels.

Let $y = (y_1,\ldots, y_n)$ be a bias vector where y_i can take on a value in $\{+1,0,-1\}$ indicating that example x_i is positively labeled, unlabeled or negatively labeled,

respectively. Let $S = (s_1,\ldots, s_n)$ be the expected score vector. We seek to find S that satisfies the following constraints: (i) S remains close to the bias vector y; and (ii) scores of two examples directly linked by an edge are not too different from each other. In learning, the algorithm computes S by first initializing S to some initial values and then iteratively updating S by performing the following steps:

1. Construct matrix $T = D^{1/2} \times W \times D^{1/2}$, where D is a diagonal matrix with $D_{ii} = d_i$ and $D_{ij} = 0$, if $j \neq i$ where $d_i = \sum_j w_{ij}$
2. $S \leftarrow y$
3. Update S till convergence: $S(t+1) \leftarrow T.S(t) + y$, where $S(t)$ is the value of S at iteration t.

In prediction, the algorithm predicts labels of nodes by thresholding the values of S.

2.3 Selecting Confident Predictions

In a bootstrapping algorithm, a subset of newly predicted labels is used to augment the labeled set to retrain the model. If incorrectly predicted labels are selected and used for training, they will affect the algorithm's accuracy. Thus, the success of bootstrapping largely depends on the algorithm's ability to select correctly predicted labels. In this subsection we will describe a method to select confidently predicted labels from score vectors after doing network propagation. Note that we consider measure of confidence as an estimation of the correctness of labels.

We call an edge of a graph *cut edge* if it connects two nodes labeled differently, i.e. +1 and −1. Let H_0 be the null hypothesis that the graph nodes are labeled independently to each other according to distribution $\{P(y=-1), P(y=+1)\}$, where $P(y=-1)$ ($P(y=+1)$) is the prior probability that an example has label $-1(+1)$ and can be computed from labeled training examples. According to *cluster assumption* that examples of the same label form clusters in graphs, a correctly labeled example should have significantly more neighbors sharing its label than observed by chance under the null hypothesis.

For an example u with predicted label y_u by using graph W, we denote the labeling confidence of u according to graph W by $CF(u,W)$ and estimate $CF(u,W)$ based on the following *cut edge weight statistic* [9]:

$$L(u,W) = \sum_{u' \in NB_u} w_{uu'} I_{uu'} \tag{1}$$

where NB_u is the set of neighbors of u ; $I_{uu'}$ is an i.i.d. binary random variable, $I_{uu'} = 1$ indicates that the edge between u and u' is cut edge. Under the null hypothesis, the probability of $I_{uu'} = 1$ is $1 - P(y = y_u)$. The expectation and variance of $L(u, W)$ under H_0 are given by:

$$\mu(u,W/H_0) = (1-P(y=y_u)) \sum_{u' \in NB_u} w_{uu'} \tag{2}$$

$$\sigma^2(u,W/H_0) = P(y=y_u)(1-P(y=y_u)) \sum_{u' \in NB_u} w_{uu'}^2 \tag{3}$$

Recall that, according to cluster assumption, for an example to be considered correctly labeled, its cut edge weight statistic L should be significantly smaller than its expected value under H_0. Under certain conditions (the size of NB_u is large enough), $L(u, W)$ can be approximately modeled by a normal distribution with mean and variance given in (3) and (4). Thus, one can estimate the left one-tailed p-value of $L(u, W)$ using a z-test, which consists of two steps: (i) computing $z_u = (1-\mu(u,W))/\sigma(u,W)$; and (ii) computing $\Phi(z_u)$, where Φ is the standard *normal cumulative distribution function*. The labeling confidence is then computed as:

$$CF(u,W) = 1-\Phi(z_u) \tag{4}$$

We consider a label assigned to example u confident if $CF(u,W)$ is larger than a predefined threshold θ_1.

Given the labeling confidence score of each graph-based learner for an example u, we then combine the scores from all the graphs to decide whether to accept the label of u for using in the next round. Intuitively, example u is considered confidently labeled if most learners agree on its confidence and its label. To measure the requirement, we compute the confidence score of u over all networks as follows:

$$CS(u) = \sum_{i=1,m} sw_i \times \mathbf{1}(u,i) \times y_u^{\,i} \tag{5}$$

where:

- sw_i is the weight of graph W_i,
- $\mathbf{1}(u,i) = 1$ if u is confidently labeled by graph W_i and $\mathbf{1}(u,i) = 0$ otherwise,
- $y_u^{\,i}$ is the label of u as predicted by graph W_i.

The graph weight sw_i specifies the contribution of the graph in computing $CS(u)$. Intuitively the graph weight should reflect the accuracy of the graph in classifying examples. Although a proper choice of sw_i would improve the accuracy by preventing the contribution of irrelevant graphs, in our experiments we used uniform weights and left the problem of choosing graph weights for future work.

Once the confidence score $CS(u)$ is computed, the predicted label of u is deemed to be correct and added to the labeled training set if $CS(u) > \theta_2$. Here θ_2 is another threshold to be defined by the user. The algorithm is summarized in Fig. 1

Function Y = CONE (V_1, …, V_m, L, U, T)

Input:
V_1, …, V_m : data sources
L: the set of labeled examples
U: the set of unlabeled example
t: the maximum number of neighbors which should be kept for each node

Output: Y: the vector of labels of all examples

Functions:
contruct_graph_from_view(.): constructs graph-representations for data sources as described in section 3.4
graph_propagate(.): predicts labels using the algorithm from [15] (see 2.2)
compute_cut_edge(.): compute cut edge weight statistic using Eq. (1).
compute_p_value(.): compute individual labeling confidence score using Eq. (4)
compute_confident_score(.): compute overall confidence score using Eq. (5)

Process:

```
for i=1:m
    W_i = contruct_graph_from_view(V_i);
end for
while(U is not empty)
    for i=1:m
        S_i = graph_propagate(W_i, L, U);
        for each u in U
                        L(u,j) = compute_cut_edge(W_i, L, U);
                        CF(u,j) = compute_p_value(L(u,j));

                    end for
    end for
    for each u in U
                    CS(u) = compute_confident_score(u, CF);
                    if(CS(u) > θ_2)
                      L = add(L, u);
                      U = remove(U, u);
                    end if
                end for
end while
score = (S_1 + S_2 + ... + S_m) / m;
for i from 1 to n
     if(score(i) > 0)        y_i = +1;
     else                    y_i = −1;
     end if
end for
```

Fig. 1. The proposed algorithm

2.4 Constructing Graphs from Data

We now describe how to construct graphs from heterogeneous data sources. Each edge connecting a pair of i-th and j-th nodes has a weight $w_{ij} \geq 0$ representing the strength of this similarity between the nodes, $w_{ij} = 0$ if there is no connection. In our experiments, we used the Pearson correlation coefficient as measure of pair-wise similarity between objects. Specifically, let x_i and x_j be the vector representations of i-th and j-th objects respectively, then w_{ij} is computed as $w_{ij} = L(x_i,x_j)$ where $L(p,q)$ is the Pearson correlation coefficient between p and q. To guarantee the sparseness of the association networks, for each node x_i we keep only K ($K << n$) vertices with the largest weights and set the others to zeros. It has been shown that keeping only small number of neighbor nodes is important not only for computational efficiency but also give more stable results of clustering and classification.

3 Experiments and Results

We evaluated the proposed method in predicting gene function from heterogeneous data sources. The problem is to assign genes to functional categories such as Gene Ontology categories (known as GO terms) (http://www.geneontology.org). We are given a set of genes, for some of which we know their functions in forms of GO terms. These genes will serve as labeled examples. The goal is to predict functions of unlabeled examples, i.e. the other genes. Gene function prediction is a multilabel classification problem, in which a gene may have more than one function. Following a common approach to multilabel learning, we solve several binary classification problems, each corresponds to assigning one specific function to the genes.

To predict gene function, one may use different types of biological data, which are heterogeneous in nature. In what follows we will describe the datasets and how we construct graphs from them.

3.1 Experimental Settings

Data. We used the yeast data set described in Barutcuoglu *et al*. [4] for experiments. This dataset contains heterogeneous data for 4524 genes, which are annotated with 105 GO terms from the *biological_process* hierarchy of Gene Ontology. GO terms are arranged in a hierarchical structure so that a gene annotated to a node in this hierarchy is also annotated to all the ancestors of this node. The dataset consists of four types of genetic data: pair-wise protein interactions, co-localization of gene products in cell, transaction factor binding sites, and microarray data. The first three types of features are binary while the last one is real-valued.

Normalization. We performed several preprocessing and normalization steps. For binary data, if an element had value b, we replaced b with $\log(1-\beta)$ if $b = 0$ and with $-\log(\beta)$ if $b = 1$ where β is the prior probability of the event $b = 1$. This transformation ensures that the similarity between two genes that share an unpopular feature is more important than the one between two genes that share a popular feature.

Evaluation method. We used 3-fold cross validation to measure the prediction accuracy of evaluated methods. We computed accuracy for each GO term and compared

them with those of other methods. For each GO term and data split, we computed the AUC score, which is the area under the ROC (Receiver Operating Characteristic) curve; then took the average value over three splits. This ROC curve expresses the trade-off between true positive rate and false positive rate at different thresholds. Therefore, it is a very suitable method to evaluate classification algorithms that return continuous scores. It is also a measure of choice where the number of positive examples is much smaller than the number of negative ones as in gene function prediction.

3.2 Results

Choosing thresholds. Thresholds θ_1 and θ_2 have important influence on the behavior of the algorithm. The higher they are, the fewer predicted labels are chosen to augment the labeled training set in each learning round and the more likely that the selected predictions are correct. Low thresholds may add more noise to the training set, while too high thresholds may lead to early stopping of the algorithm. We varied values of θ_1 and θ_2 and measured the AUCs on held-out test sets. In what follows we report the best results corresponding $\theta_1 = 0.8$ and $\theta_2 = 0.5$.

Comparison with LGC. We compared the proposed method with algorithm LGC (learning with Local and Global Consistency) described in [15]. We used the conjugate gradient based implementation of solver provided by Mostafavi *et al*. [7] with all settings mentioned in section D of [7]. In the first experiment, LGC was run on individual graphs, each represents a single data source, and the corresponding AUC score was computed. Fig. 2 shows the average AUC scores over 105 GO terms of LGC on each individual graph and that of our method. As shown, the proposed method performed the best and achieved average AUC of 0.87 over 105 GO terms.

In the next experiment, we ran LGC on the combined graph, which is the sum of the four single graphs. Such combination has been reported in previous work [7,16] to have improved prediction accuracy over individual graphs. Fig. 3a plots the AUC estimations of LGC on the combined graph (LGC-UNI) and those of our method, for 105 GO terms. Each filled circle in Fig. 3 represents a GO term. The x-axis plots the AUC of LGC and the y-axis plots the AUC of our method. A circle above the diagonal represents the improvement of our method in terms of AUC for the corresponding GO term. As shown, our method showed improved AUC for 102 out of 105 GO terms. LGC performed better or equally for only 3 GO terms.

Comparison with SW. We also compared our method with SW [8,7] – a state-of-the-art gene prediction method. Given a collection of individual networks, SW constructs a composite network, which is a weighted linear combination of individual networks. The composite network is constructed by solving a constrained linear regression problem where weights are simultaneously optimized on a group of GO terms. The method has shown leading performance in predicting gene function for several species such as yeast, human, fly and mouse and at the same time is very fast.

Fig. 3.b compares the AUC scores of the proposed method with those of SW. Over 105 GO terms, SW achieved average AUC of 0.837, which is the second best AUC score among all the tested method (after the proposed one). Again, our method showed higher average AUC and scored better than SW on 79 GO terms. The difference was significant according to Wilcoxon signed rank test (p-value < $1\text{x}10^{-3}$).

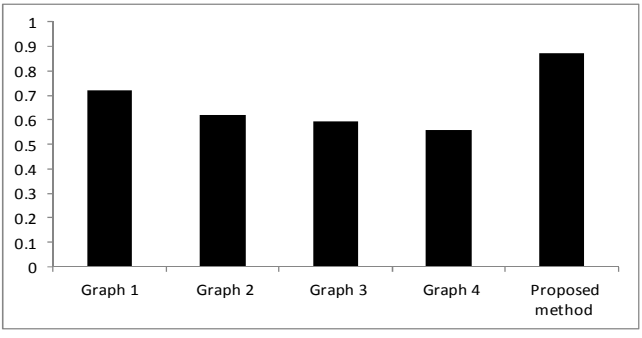


Fig. 2. Average AUCs of LGC on individual graphs and average AUC of the proposed method

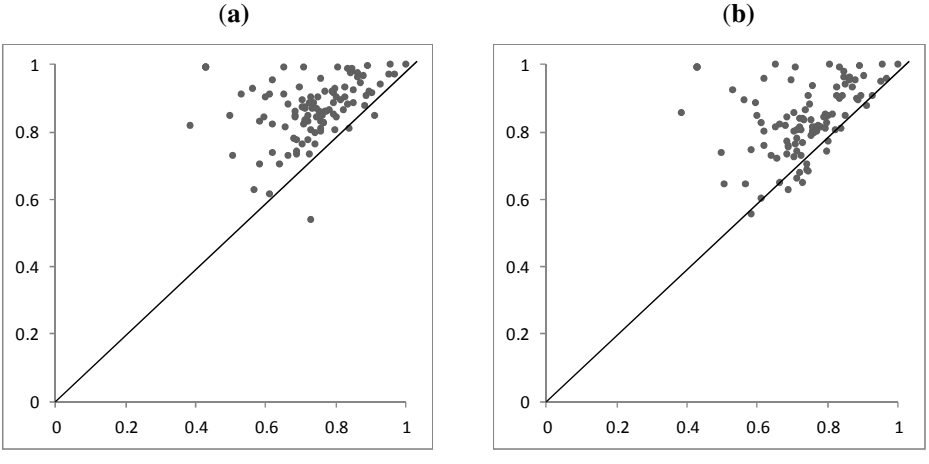


Fig. 3. Comparison the AUCs of the proposed method with LGC-UNI (**a**) and SW (**b**). Each point represents a GO term, plotting AUC for the proposed method on the y-axis and AUC for LGC-UNI (**a**) and SW (**b**) on the x-axis. Points above the diagonal represents GO terms for which the proposed methods achieved higher AUC scores.

4 Conclusion

We have presented a method for learning from multiple heterogeneous data sources. By combining bootstrapping with graph-based learning, the method inherits the strengths of both approaches. The key factor for the success of bootstrapping lies on how to correctly select highly confident predictions to add to the labeled set. Our method addresses this issue by using a measure called cut edge weight statistic and extending its use to multiple graphs. Experiments in predicting gene function show promising results where our method outperforms two other methods including a standard graph-based method and a leading gene function prediction system.

Acknowledgments. This work was supported by the National Foundation for Science and Technology Development (NAFOSTED) of Vietnam.

References

1. Abney, S.: Bootstrapping. In: Proceedings of ACL 2002, pp. 360–367 (2002)
2. Argyriou, A., Herbster, M., Pontil, M.: Combining graph Laplacians for semi- supervised learning. In: Advances in Neural Information Processing Systems18, NIPS 18, MIT Press, Cambridge (2006)
3. Balcan, M.-F., Blum, A., Yang, K.: Co-training and expansion: Towards bridging theory and practice. In: NIPS 17, pp. 89–96 (2005)
4. Barutcuoglu, Z., Schapire, R., Troyanskaya, O.: Hierarchical multi-label prediction of gene function. Bioinformatics 22(7), 830–836 (2006)
5. Blum, A., Mitchel, T.: Combining labeled and unlabeled data with co-training. In: Proc. Workshop on on Computational Learning Theory, COLT 1998, pp. 92–100 (1998)
6. Chapelle, O., Scholkopf, B., Zien, A. (eds.): SemiSupervised Learning. MIT Press (2006)
7. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., Morris, Q.: GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biology 9(suppl. 1), S4 (2008)
8. Mostafavi, S., Morris, Q.: Fast integration of heterogeneous data sources for predicting gene function with limited annotation. Bioinformatics 26(14), 1759–1765 (2010)
9. Muhlenbach, F., Lallich, S., Zighed, D.A.: Identifying and handling mislabeled instances. Journal of Intelligent Information Systems 22(1), 89–109 (2004)
10. Pavlidis, P., Weston, J., Cai, J., Grundy, W.N.: Gene functional classification from heterogeneous data. In: Proceedings of RECOMB 2001, pp. 249–255 (2001)
11. Tang, W., Lu, Z., Dhillon, I.S.: Clustering with multiple graphs. In: ICDM 2009, pp. 1016–1021 (2009)
12. Tsuda, K., Shin, H., SchÄolkopf, B.: Fast protein classification with multiple networks. Bioinformatics 21, ii59–ii65 (2005)
13. Wang, W., Zhou, Z.H.: A new analysis of co-training. In: Proceedings of International Conference on Machine Learning, ICML 2010, pp. 1135–1142 (2010)
14. Zhang, M.L., Zhou, Z.H.: CoTrade: Confident co-training with data editing. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 99, 1–15 (2011)
15. Zhou, D., Bousquet, O., Lal, T., Weston, J., Scholkopf, B.: Learning with local and global consistency. In: Advances in Neural Information Processing Systems, vol. 16. MIT Press, Cambridge (2004)
16. Zhou, D., Burges, C.J.C.: Spectral clustering and transductive learning with multiple views. In: Proceedings of the 24th International Conference on Machine Learning, ICML 2007, pp. 1159–1166 (2007)
17. Zhou, Z.-H., Li, M.: Tri-training: Exploiting unlabeled data using three classifiers. IEEE Transactions on Knowledge and Data Engineering 17(11), 1529–1541 (2005)
18. Zhu, X.: Semi-supervised learning literature survey. Technical Report 1530, Department of Computer Science, University of Wisconsin at Madison (2008)

Author Index