

IWIN2013



International Workshop on Informatics

~~AAAAA~~ Proceedings of
International Workshop on
Informatics
September 1-4, 2013
Stockholm, Sweden



Sponsored by Informatics Society

Publication office:

Informatics Laboratory

3-41, Tsujimachi, Kitaku, Nagoya 462-0032, Japan

Publisher:

Tadanori Mizuno, President of Informatics Society

ISBN:

978-4-902523-34-8

Printed in Japan

Table of Contents

Session 1: Network 1

(Chair: Yoh Shiraishi) (10:00 - 12:30, Sept. 2)

(1)	A Proposal of P2P Content Delivery System for Supporting Streaming Applications	3
	Takanori Kashiwagi, Jun Sawamoto, Hiroyuki Sato, Yuji Wada, Norihisa Segawa, and Eiji Sugino	
(2)	Reducing Probe Data in Telematics Services Using Space and Time Models	9
	Ryozo Kiyohara, Hirohito kakizawa, Shinji Kitagami, Yoshiaki Terashima, and Masashi Saito	
(3)	Reactive Load Balancing During Failure State in IP Fast Reroute Schemes	15
	Kazuki Imura, and Takuya Yoshihiro	
(4)	A DTN Routing Scheme Based on Publish/Subscribe Model	23
	Ryosuke Abe, Yoshitaka Nakamura, and Osamu Takahashi	
(5)	Sales and marketing support for BtoB using Web form	31
	Hiroshi Horikawa	

Session 2: Education and Business

(Chair: Tomoo Inoue) (14:00 - 16:30, Sept. 2)

- (6) Proposal and Evaluation of Collaborative Attribute Method in Text
Recommender Systems for E-Learners 41
Yuji Wada, Takuya Segawa, Jun Sawamoto and Hiroyuki Sato
- (7) Application of a Lump-sum Update Method to Distributed Database 49
Tsukasa Kudo, Yui Takeda, Masahiko Ishino, Kenji Saotome and
Nobuhiro Kataoka
- (8) Comparative analysis of cognition and memorization during learning
using simple electroencephalographs 57
Koji Yoshida, Humiyasu Hirai and Isao Miyaji
- (9) Sugoroku Game Interactions with Twitter 65
Jun Munemori, Kanako Morimoto and Junko Itou
- (10) Extension Mechanism for Integrating New Technology Elements into
Viewpoint based Enterprise Architecture Framework 71
Akira Tanaka and Osamu Takahashi

Session 3: Communication

(Chair: Tomoya Kitani) (16:45 - 19:15, Sept. 2)

- (11) Disaster-Relief Training System Using Electronic Triage with Voice
Input 83
Misaki Hagino, Yoshiaki Ando and Ken-ichi Okada
- (12) A Study of Disaster Library System with a Field Agent to Learn a Sequence of
Great Disasters 91
Taizo Miyachi, Gulbanu Buribayeva, Saiko Iga and Takashi Furuhata
- (13) A System to Help Creation of Original Recipes by Recommending Additional
Foodstuffs and Reference Recipes 99
Mana Tanaka, Etsuko Inoue, Takuya Yoshihiro, Masaru Nakagawa
- (14) An Estimation Method of Consumption Calories in Activities of Daily Living
based on METs values 107
Yoshitaka Nakamura, Yoshiki Matsubayashi, Yoh Shiraishi and Osamu
Takahashi
- (15) User's Communication Behavior in the Pseudo Same-room Videoconferencing
System 115
Mamoun Nawahdah and Tomoo Inoue

Session 4: System, Software and Algorithm

(Chair: Takuya Yoshihiro) (10:00 - 12:30, Sept. 3)

- (16) Variable Coverage: A Metric to Evaluate the Exhaustiveness for Program Specifications Based on DbC 125
Yuko Muto, Yukihiro Sasaki, Takafumi Ohta, Kozo Okano,
Shinji Kusumoto and Kazuki Yoshioka
- (17) Development of a Learning Tool for Computer Programming Using a Robot Remotely Controlled by a PC through Bluetooth Wireless Communication 135
Toshihiro Shikama
- (18) A Design Method of Optimal H2 Integral Servo Problem 143
Noriyuki Komine, Masakatsu Nishigaki, Kunihiro Yamada and
Tadanori Mizuno
- (19) Subgrid Search algorithm for solving Hitorinishitekure 151
Shohei Okuyama and Naoya Chujo
- (20) Verification of a Control Program for a Line Tracing Robot using UPPAAL Considering General Aspects 157
Toshifusa Sekizawa, Kozo Okano, Ayako Ogawa and Shinji Kusumoto

Keynote Speeche (14:00 - 15:00, Sept. 3)

- (I) Research Activities in DOCOMO ~ Toward Smart Life Partner ~167
Hiroshi Inamura

Session 5: Network 2

(Chair: Yoshitaka Nakamura) (15:30 - 18:00, Sept. 3)

- (21) Improving of Terminal-Independent Handover Method with SIP
Mobility 181
Yoshio Oda, Yoshitaka Nakamura and Osamu Takahashi
- (22) Improvement of The Communication Stability for Wireless M2M Router
System 189
Kazuaki Honda and Osamu Takahashi
- (23) Method to extract Genre or Character from tags of illustration via
Pixiv 195
Eiichi Takebuchi, Yasuhiro Yamada, Akira Hattori and Haruo Hayami
- (24) A Method for Information Delivery in VANET with Scheduled Routes of
Vehicles 201
Masato Nakamura, Tomoya Kitani, Weihua Sun, Naoki Shibata,
Keiichi Yasumoto and Minoru Ito

Panel Discussion: Social Computing

(11:00 - 12:30, Sept. 3)

Chair

- Prof. Yoshimi Teshigawara, Tokyo Denki University

Panelists

- Prof. Tadao Obana, The University of Electro-Communications
- Prof. Teruo Higashino, Osaka University
- Dr. Hiroshi Inamura, Research Laboratories, NTT DOCOMO, Inc.

A Message from the General Chair



It is my great pleasure to welcome all of you to Stockholm, Sweden, for the Seventh International Workshop on Informatics (IWIN 2013). This workshop is sponsored by the Informatics Society. The first workshop was held in Napoli, Italy in September 2007. Since then, workshops from the second to the sixth were held annually: in Wien, Austria in September 2008; in Hawaii, USA in September 2009; in Edinburgh, Scotland, UK in September 2010; in Venice, Italia in September 2011; and in Chamonix, France in September 2012.

In IWIN 2013, 24 papers have been accepted. They cover a wide range of area in informatics. Based on the papers, 5 technical sessions have been organized in a single track format: Network1; Education and Business; Communication; System, Software and Algorithm; Network2. In addition, IWIN 2013 has an invited session from Dr. Hiroshi Inamura who is an executive research engineer of Research Laboratories, NTT DOCOMO, Inc. And, we really appreciate the participation of the invited speakers in this workshop.

By the way, Stockholm is a famous place for the Nobel Prize, and it is widely known that Professor Shinya Yamanaka was awarded last year, 2012. So, we will hold our banquet with the Nobel Prize cuisine. We hope this conference will be participant's motivations for the great research achievement.

And, I would like to thank all of participants and contributors who made the workshop possible. It is indeed an honor to work with a large group of professionals around the world for making the workshop a great success.

We are looking forward to seeing you all in the workshop. We hope you all will experience a great and enjoyable meeting in Stockholm, Sweden.

A handwritten signature in black ink that reads "Tsukasa Kudo". The signature is written in a cursive, flowing style.

Tsukasa Kudo
General Chair

The International Workshop on Informatics 2013

Organizing Committee

General Chair

Tsukasa Kudo (Shizuoka Institute of Science and Technology, Japan)

Steering Committee

Toru Hasegawa (Osaka University, Japan)

Teruo Higashino (Osaka University, Japan)

Tadanori Mizuno (Aichi Institute of Technology, Japan)

Jun Munemori (Wakayama University, Japan)

Yuko Murayama (Iwate Prefectural University, Japan)

Ken-ichi Okada (Keio University, Japan)

Norio Shiratori (Tohoku University, Japan)

Osamu Takahashi (Future University-Hakodate, Japan)

Finance Chair

Tomoya Kitani (Shizuoka University, Japan)

Program Committee Chair

Kozo Okano (Osaka University, Japan)

Program Committee

Behzad Bordbar (University of Birmingham, UK)

Naoya Chujo (Aichi Institute of Technology, Japan)

Satoru Fujii (Matsue College of Technology, Japan)

Teruyuki Hasegawa (KDDI R&D Laboratories, Japan)

Haruo Hayami (Kanagawa Institute of Technology, Japan)

Takaaki Hishida (Aichi Institute of Technology, Japan)

Nobutsugu Fujino (Fujitsu Laboratories, Japan)

Tomoo Inoue (University of Tsukuba, Japan)

Masahiko Ishino (Fukui University of Technology, Japan)

Yoshinobu Kawabe (Aichi Institute of Technology, Japan)

Gen Kitagata (Tohoku University, Japan)

Tomoya Kitani (Shizuoka University, Japan)

Teruo Higashino (Osaka University, Japan)

Hiroshi Mineno (Shizuoka University, Japan)

Shinichiro Mori (Fujitsu Laboratories, Japan)

Jun Munemori (Wakayama University, Japan)

Yoshitaka Nakamura (Future University-Hakodate, Japan)

Yuji Wada (Tokyo Denki University, Japan)

Masakatsu Nishigaki (Shizuoka University, Japan)

Yoshia Saito (Iwate Prefectural University, Japan)

Fumiaki Sato (Toho University, Japan)

Jun Sawamoto (Iwate Prefectural University, Japan)

Hiroshi Shigeno (Keio University, Japan)

Toshihiro Shikama (Fukui University of Technology, Japan)

Yoh Shiraishi (Future University-Hakodate, Japan)

Hideyuki Takahashi (Tohoku University, Japan)

Osamu Takahashi (Future University-Hakodate, Japan)

Yoshiaki Terashima

(Mitsubishi Electric Corporation, Japan)

Takaaki Umedu (Shiga University, Japan)

Hirozumi Yamaguchi (Osaka University, Japan)

Koji Yoshida (Shonan Institute of Technology, Japan)

Takuya Yoshihiro (Wakayama University, Japan)

Takaya Yuizono

(Japan Advanced Institute of Science and Technology, Japan)

Session 1:

Network 1

(Chair: Yoh Shiraishi)

A Proposal of P2P Content Delivery System for Supporting Streaming Applications

Takanori Kashiwagi^{*}, Jun Sawamoto^{*}, Hiroyuki Sato^{*}, Norihisa Segawa^{*}, Eiji Sugino^{*}, Yuji Wada^{**}

^{*}Faculty of Software and Information Science, Iwate Prefectural University
152-52 Sugo, Takizawa, Iwate 020-0193 JAPAN
{sawamoto, sato_h}@iwate-pu.ac.jp

^{**}Department of Information Environment, Tokyo Denki University
2-1200 Muzaigakuendai, Inzai-shi, Chiba-ken 270-1382 JAPAN
yujiwada@sie.dendai.ac.jp

Abstract -Streaming large files such as video and audio contents from the internet has become an increasingly common practice with users and content providers. Content delivery presents serious challenge for content providers, with the increased cost of hosting and transmitting large video files, the existing client server system is experiencing problems. The high server load incurred by the client model is costing hosts considerable resources. Peer to Peer (P2P) technology alleviates some of these problems by distributing transfer work among multiple hosts (peers). P2P works by sending and receiving data directly with other peers that are participating in the network. It distributes resources and load across the network. This can solve the problem of the client server system resource overload. The purpose of this research is to propose a method which is suitable for streaming using P2P and solve the problem of client server system's resource overload. We aim to realize stable video streaming, low latency playback, and reduction of the number of breaks due to buffering protocol.

Keywords: Content delivery, Streaming, Peer to Peer network, BitTorrent, BiToS.

1 INTRODUCTION

The video and audio content delivery service using the internet, such as YouTube [1] and NicoNico Douga [2], has become an increasingly common practice, and it is capturing the attention from broad directions, such as political use and commercial use, etc. Moreover, by the development of broadband service and improvement of terminal performance of individual use, it is expected that the video and audio content as a medium for disseminating information continues to grow. In the prediction and investigation of Cisco [3], it is expected that two-thirds of the world's mobile data traffic will be video by 2017. Mobile video will increase 16-fold between 2012 and 2017, accounting for over 66 percent of total mobile data traffic by the end of 2017. As streaming large files such as video and audio content from the internet has become an increasingly common practice with users and content providers, the content delivery presents serious challenge for content providers, with the increased cost of hosting and transmitting large video files, the existing client server system is experiencing problems. The high server load incurred by the server-client model is costing hosts considerable resources.

Peer to Peer (P2P) technology alleviates some of these problems by distributing transfer work among multiple hosts (peers). BitTorrent [4] is one of the most popular P2P protocols. File transfer operates by splitting the file into many pieces. Peers transfer the pieces out of order in a distributed fashion then re-assemble the original file. The order of the pieces transferred is determined by the RarestFirst algorithm [5][6]. However, it is bad for streaming because pieces are transferred out of order and it is hard to predict the next piece. BiToS (BitTorrent Streaming) [7] was proposed to solve the streaming P2P problems of BitTorrent. This allowed somewhat smoother playback, but there were still delays or pauses (breaks). And some new methods to shorten breaks' time and reduce the number of times of breaks are called for.

We propose a method which is suitable for streaming using P2P. The emphasis must be placed on reduction of the number of breaks in playback. To this end, we must do something different if there is a gap in download pieces between current playback position and the next available piece. Improved peer and piece selection methods, such as special priority for pieces near playback position may hopefully alleviate the problems with BiToS and RarestFirst algorithm. Specifically, if the piece closest to the playback position is not yet downloaded then the proposed method will set an emergency priority. Within the high priority group we must request missing pieces from the peer with the fastest connection. In order to verify the proposed method's effectiveness when compared to the established methods of RarestFirst and BiToS, we performed simulations and experiments.

The rest of this paper is organized as follows: In section 2, we describe detailed algorithm of BitTorrent and BiToS. In section 3, we present our proposed solution for better peer and piece selection. In section 4, details of the implementation on software simulator is described. In section 5, we report experimental evaluation of our proposed method. Finally, the paper is concluded in section 6.

2 BITTORRENT AND BITOS

BitTorrent is one of the most popular P2P protocols. Holding, sending, and receiving of all content are performed by only the peers. The tracker manages information about peers in a swarm; it coordinates initial connections and keeps a table of connected hosts and the download/upload statistics of each peer (Fig.1).

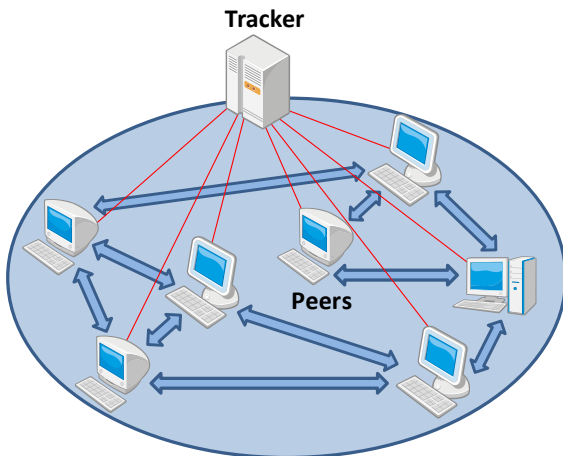


Figure 1: Network configuration of BitTorrent.

As shown in Fig.2, BitTorrent uses swarming techniques in which the torrent file (the content that is distributed), is split in pieces. A user who wants to upload a file first acts as a seed and distributes content information through BitTorrent nodes. Peers (leecher) can simultaneously download pieces from other peers. While the peer is downloading pieces of the file, it uploads the pieces that it has already acquired to its peers. Each time the peer has a new piece, it advertises this information to its peer set (the peers that the peer is connect to).

Peers transfer the pieces out of order in a distributed fashion then re-assemble the original file. This distributed method is suitable for large-capacity content delivery.

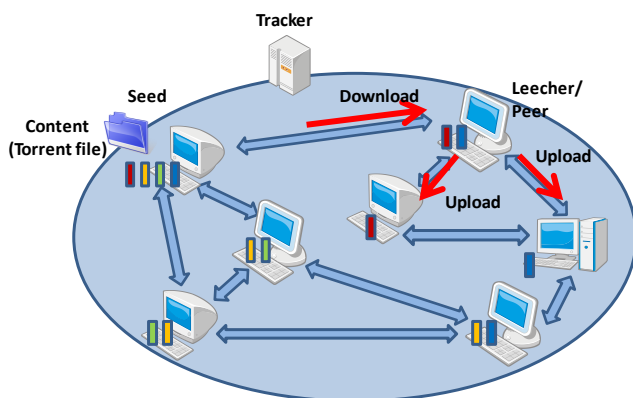


Figure 2: File transfer operates by splitting the file into many pieces

The order of the pieces transferred is determined by the RarestFirst algorithm. This algorithm tells peers to send the least common pieces amongst the swarm first, causing convergence faster. RarestFirst transfer makes P2P very efficient when compared to the random out of order method. However, it is bad for streaming because pieces are transferred out of order and it is hard to predict the next piece. Streaming requires in-order transfer for smooth playback. The method proposed in this paper aims to provide more predictable transfer to allow for smooth playback.

BiToS was a previous attempt to solve the streaming P2P problems (Fig.3). It was a research to reduce the number of breaks when streaming using BitTorrent. The BiToS method changed from RarestFirst so that pieces near deadline mark have higher priority than later pieces. This allowed somewhat smoother playback, but there were still pauses. BiToS method works by assigning a priority to two groups of pieces. If the probability of selecting a piece from the high priority group is “ p ” then low priority group probability is “ $1-p$ ”. The parameter “ p ” represents the balance between the immediate need for a piece and the future need. Within each priority group we simply use RarestFirst method.

Currently downloading pieces in high priority group and low priority group are moved to the group of received pieces after they are downloaded. If a piece cannot meet its playback deadline, then it will not be asked to be downloaded (or its download can be aborted) and will be marked Missed. A peer at any given time can have at maximum a fixed number of currently downloading pieces. The number of pieces (cardinality) of the higher priority group remains fixed. Using BiToS, we receive pieces closer to the playback position sooner. This is more suitable for content delivery than pure RarestFirst method.

However within each group the RarestFirst method is still used, so there may be breaks if the priority group has not rare pieces close to the playback position. This means pieces are still sent out of order within each priority group. This causes gaps in playback when the playback position reaches a missed piece.

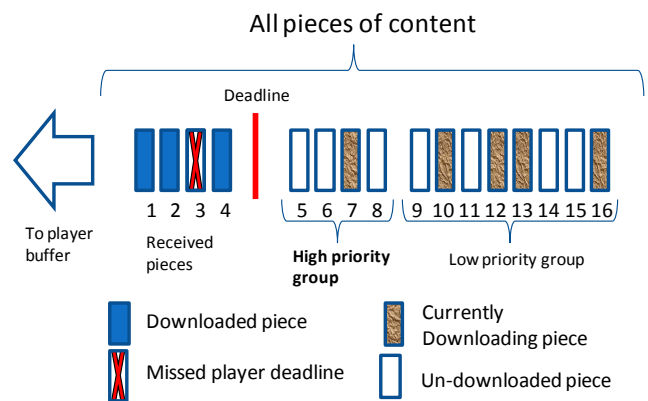
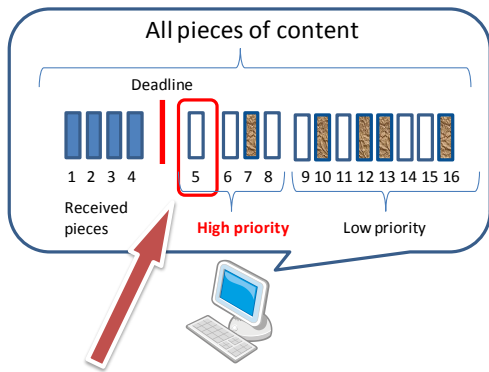


Figure 3: Outline of BiToS piece selection method.

3 PROPOSED SOLUTION

To propose a method which is suitable for streaming using P2P, emphasis must be placed on the reduction of the number of breaks in playback. To this end, we must do something different if there is a gap in download pieces between our deadline position and the next available piece. Here, the deadline is the time limit after that, the received piece is not useful and will be discarded.



If the piece closer to the deadline position is not yet downloaded then set an emergency priority

Figure 4: Introduction of emergency priority.

Improved peer and piece selection methods, such as special priority for pieces near deadline position may hopefully alleviate the problems with BiToS and RarestFirst. Specifically, if the piece closest to the deadline position is not yet downloaded then the proposed method will set an emergency priority (Fig. 4). Within the high priority group we must request emergent pieces from the peer with the fastest connection (Fig. 5).

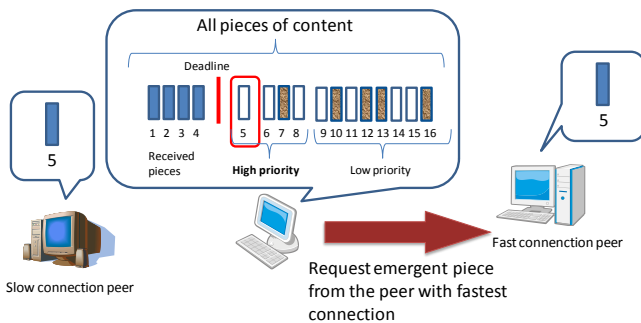


Figure 5: Request emergent pieces from the peer with the fastest connection.

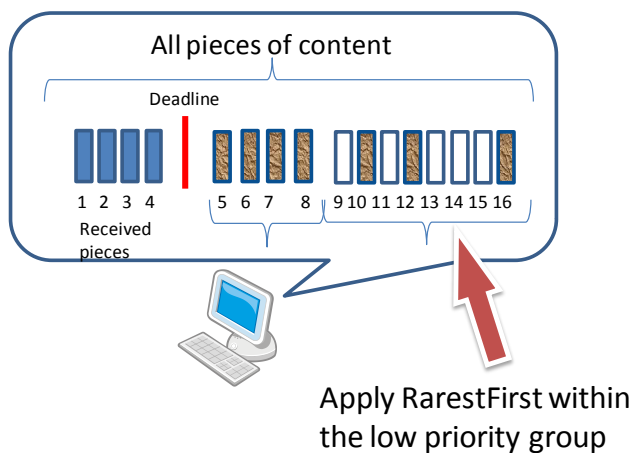


Figure 6: Enough buffered content then download pieces from a lower priority group.

If there is enough buffered content then the new method may download pieces from a lower priority group using simple RarestFirst (Fig. 6). Thus it is still possible to contribute to the distribution of rare pieces on low priority groups and improve convergence speed.

The proposed method solves the problem of BiToS where pieces close to playback position are not always chosen. This leads to a more stable delivery and smooth playback.

4 IMPLEMENTATION ON A SOFTWARE SIMULATOR

In order to verify the proposed method's effectiveness when compared to the established methods of RarestFirst and BiToS, it is necessary to perform simulations and experiments. One such proposed experiment is to provide a peer that implements each method on a software simulator. We used General Purpose Simulator for P2P network (GPS) [8] which is capable of simulating BitTorrent algorithm.

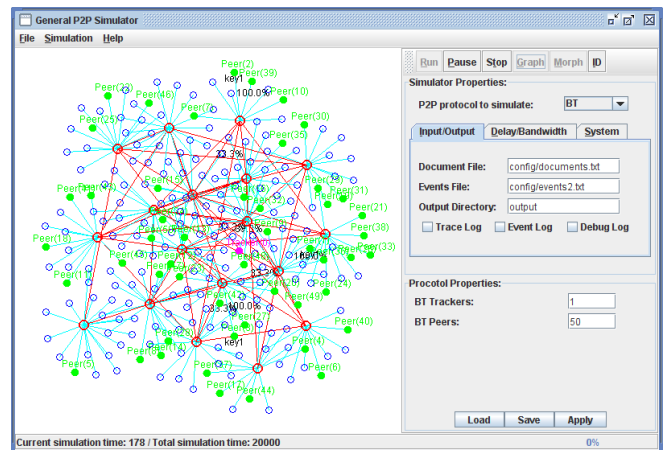


Figure 7: Display image of the simulation by General Purpose Simulator for P2P network.

As for the software structure of GPS, various search protocols such as Chord [9], CAN [10], etc. are located on top of the physical network layer at the bottom of the structure. The layer of P2P algorithms come on the search layer. Some Hybrid P2P algorithms including BitTorrent exist in the same layer as the search layer, because they don't use provided general search protocols like Chord etc. but they mostly implement original search protocols using the server systems.

The methods of previous works and our proposed method are implemented on top of the P2P algorithms layer, and they can be switched according to the experimental situation. However, it is not possible to make peers who adopt different methods on the same network at present.

Moreover, in the operation of the various methods, since it is necessary to acquire the information of the playback position, and to measure the number of times of breaks and duration and frequency of breaks, which is the evaluation indices, we added virtual video player part on top of the P2P algorithm layer.

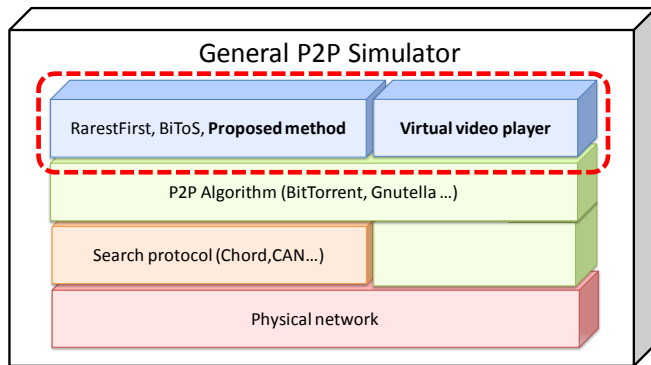


Figure 8: Software structure of GPS.

5 EXPERIMENTS

The peer and piece selection method proposed by this research, RarestFirst and BiToS are compared by measuring evaluation indices such as the number of breaks and the total duration of breaks under various video download conditions.

5.1 Outline of the experiment

First, the peer who had all the pieces (original content) is generated on the simulator. Then a peer who does not have any piece participates one at a time to the network with certain interval and starts content downloading. Playback is started when the head piece of the content is downloaded at the peer. All the peers continue remaining in the network until the last peer completes the download. All the peers who participated to the network complete the download of whole content and finish the playback then the simulation stops.

The transmission speed of peers are classified into two types such as high speed and low speed, and randomly assigned to each peer. In the communication between low-speed peers, bandwidth is set to 5Mbps, between a high-speed peer and a low-speed peer 10Mbps, and between high-speed peers 15Mbps.

Simulations are iterated 10 times for each method respectively, and the results are compared on the average basis.

5.2 Contents and parameters used for the simulation

The details of parameters used for the simulation are shown in Table 1. The content sizes are two kinds, 128 MB and 256 MB.

The size per one piece, in consideration of the size length used widely when dividing a file by BitTorrent, is set as 1 MB. Even if the content size is the same, the playback time differs according to the content quality, high and low image quality. We experiment two cases of playback time, i.e., 0.5 seconds and 4 seconds per one piece, supposing two content qualities.

Table 1: Details of the contents and parameters used for the simulation

Content size (Mbyte)	128MB		256MB	
Size of a piece (Mbyte)	1MB			
Playback time per a piece (sec)	4	0.5	4	0.5
Number of peers	50			
Participating interval of new peers (sec)	60			
Ratio of the high priority group (%)	5			
Probability of selecting a piece from high priority group (%)	90			

5.3 Experimental results and evaluation

5.3.1 Content size 128MB, 4 seconds of playback time per one piece

The experimental result in case of content size is 128 MB and the playback time per one piece is 4 seconds is shown here. Fig.9 is a graph of the total of the duration of breaks in average at each peer during the playback by each method. The total duration of breaks at the peer which completed download earlier is large and decreases as the number of peers increases for all methods. This is because when few peers are in the network, the number of downloadable peers is small, but it increases as more peers participate to the network and the feature of P2P algorithm that a download speed rises using a communication line effectively as the number of peers increase is shown here. From the graph, significant difference is not seen as a whole by each method, but when the average was taken for each method, it turned out that the total of the duration of breaks in average is the shortest in our proposed method than RarestFirst or BiToS.

Fig.10 shows the frequency distribution of the duration of breaks at each peer for each method. In our proposed method, many peers have shorter duration of breaks compared to other methods and it could be assumed that many peers have achieved shorter download time of the content as a whole.

On the other hand, about the number of times of breaks, as shown in Fig.11, no method is stable and no significant difference is seen in average here.

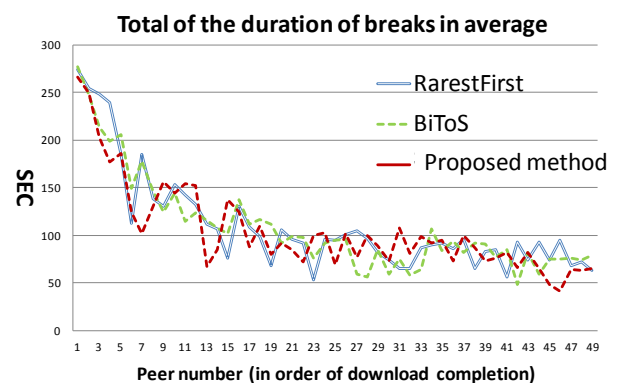


Figure 9: Total of the duration of breaks in average at each peer (Content size 128MB, 4 seconds of playback time per one piece).

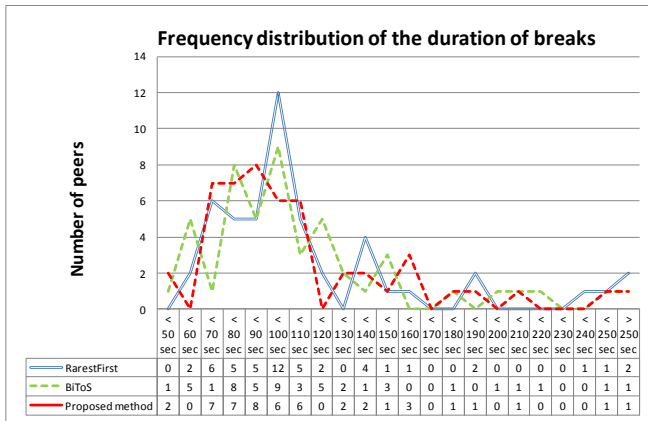


Figure 10: Frequency distribution of the duration of breaks at each peer in average (Content size 128MB, 4 seconds of playback time per one piece).

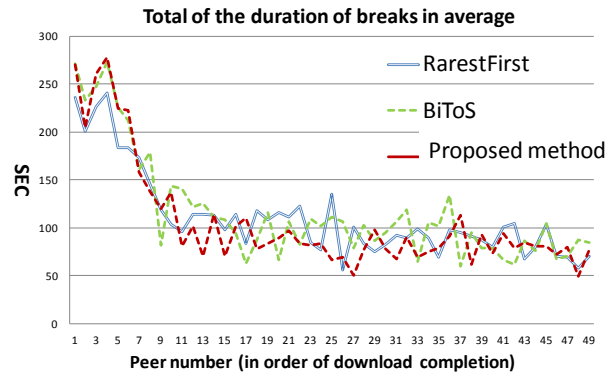


Figure 12: Total of the duration of breaks in average at each peer (Content size 128MB, 0.5 seconds of playback time per one piece).

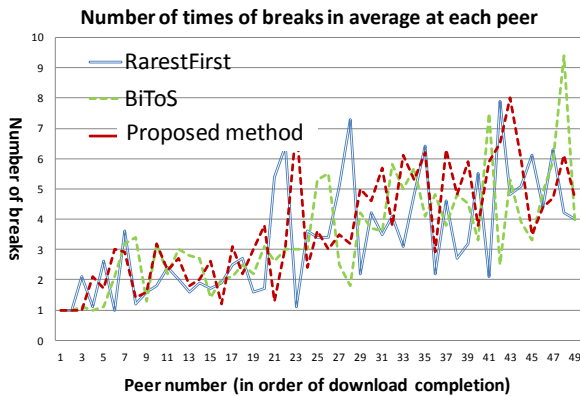


Figure 11: Number of times of breaks in average (Content size 128MB, 4 seconds of playback time per one piece).

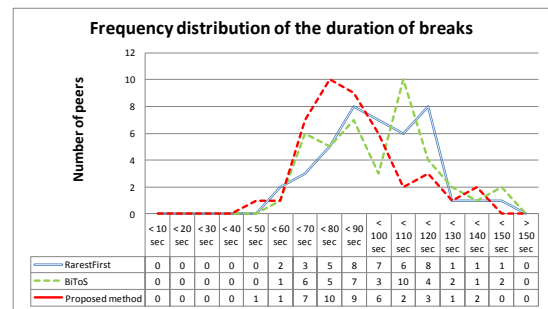


Figure 13: Frequency distribution of the duration of breaks at each peer in average (Content size 128MB, 0.5 seconds of playback time per one piece).

5.3.2 Content size 128MB, 0.5 seconds of playback time per one piece

The experimental result in case of content size is 128 MB and the playback time per one piece is 0.5 seconds is shown here. From the graph of Fig.12, the total duration of breaks in average at each peer during the playback shows similar trend as the case of 4 seconds of playback time per one piece, and it turned out that the total of the duration of breaks is the shortest in average in our proposed method.

The frequency distribution of the duration of breaks at each peer in Fig. 13 shows that the number of peers of less than 100 seconds of duration of breaks is the largest in our method.

On the other hand, about the number of times of breaks, no big difference is seen among methods just like the case of 4 seconds of playback time per one piece.

5.3.3 Content size 256MB, 4 seconds of playback time per one piece

The experimental result in case of content size is 256 MB and the playback time per one piece is 4 seconds is discussed here. The proposed method has shown poor performance here and the total duration of breaks in average at each peer is the largest as shown in Fig.14.

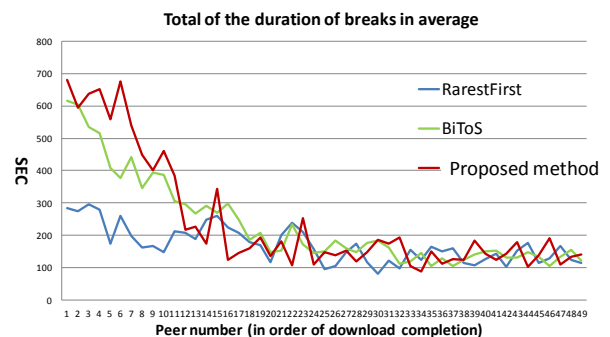


Figure 14: Total of the duration of breaks in average at each peer (Content size 256MB, 4 seconds of playback time per one piece).

In terms of the frequency distribution of the duration of breaks at each peer, the number of peers of less than 200 seconds of duration of breaks is the largest in our method. The number of breaks in average is the smallest in our method, but no significant difference is seen by methods.

5.3.4 Content size 256MB, 0.5 seconds of playback time per one piece

The experimental result in case of content size is 256 MB and the playback time per one piece is 0.5 seconds is discussed here. Here also the proposed method performed poorly in terms of total duration of breaks in average. The frequency distribution of the duration of breaks shows the distribution is high in the area of 130-200 seconds and over 250 seconds area in all methods. The number of breaks in average is the smallest in our method, but no significant difference is seen by methods here also.

5.4 Consideration

In case of content size 128MB, in both cases of 4 and 0.5 seconds of playback time per one piece, the number of times of breaks is rather small in all peers and no significant difference was seen by each method. It is considered that since the communication with sufficient bandwidth is secured by any method because the size of the content is small enough for the environment with assumed number of peers and line speed. On the other hand, there is less number of times of breaks in case the playback time per one piece is 4 seconds rather than the case of 0.5 second. This indicates that long playback contents with low image quality have less frequent breaks. About the duration of breaks, in both cases of 4 and 0.5 seconds of playback time per one piece, the average duration of breaks is the smallest by our proposed method. In many peers, average duration of breaks distributes between 50 to 120 seconds. In case of 0.5, the duration came between 50 to 100 in most of peers by our method, and our proposed method performed better than other methods.

In case of content size is 256MB, in both cases of 4 and 0.5 seconds of playback time per one piece, average number of times of breaks is smallest by our method, but no significant difference is seen among methods. This is because the content size is rather large and pieces are too many for the assumed environment in this case. For the duration of breaks, in both cases of 4 and 0.5 seconds of playback time per one piece, the average duration of breaks is the largest by our method. And from the frequency distribution of the duration of breaks, distribution of short breaks is almost same by all methods, but breaks of long duration are seen in many peers by our method. This is considered that when the system downloads pieces with emergency priorities, download requests from other peers also swarm about a certain peer and causes a long waiting time for the download request.

6 CONCLUSION

The purpose of this research is to propose a method which is suitable for video streaming using P2P while solving the problem of client server system resource overload in the content delivery market. The research has proposed a new method of peer and piece selection in a P2P streaming environment using BitTorrent. The proposed simulations examine the effectiveness of the new methods for improving on the established BiToS and RarestFirst methods. It is the research's sincerest hope that the proposed method alleviates some of the current challenges facing streaming content delivery.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number 24500122.

REFERENCES

- [1] YouTube : <http://www.youtube.com/>
- [2] NicoNico Douga : <http://www.nicovideo.jp/>
- [3] Cisco®, Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2012–2017, (2013).
- [4] B. Cohen, Incentives build robustness in bittorrent, In 1st Workshop on the Economics of Peer-2-Peer Systems, Berkley, CA, June 5-6 2003.
- [5] BitTorrent Specifications. <https://wiki.theory.org/BitTorrentSpecification> .
- [6] Arnaud Legout, G. UrvoyKeller, and P. Michiardi, Rarest First and Choke Algorithms Are Enough, IMC '06 Proceedings of the 6th ACM SIGCOMM conference on Internet measurement , pp.203–216, 2006.
- [7] A. Vlavianos, M. Iliofotou, and M. Faloutsos, BiToS: Enhancing Bittorrent for Supporting Streaming Applications, INFOCOM 2006, Proc. of 25th IEEE International Conference on Computer Communications, pp.1–6, (2006).
- [8] Weishuai Yang, Nael Abu-Ghazaleh, “GPS: A General Peer-to-Peer Simulator and its Use for Modeling BitTorrent” , Proceedings of 13th Annual Meeting of the IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS '05), (2005).
- [9] Stoica, I., et al. Chord: A scalable peer-to-peer lookup service for Internet applications. In Proceedings of ACM SIGCOMM, Volume 31 Issue 4, pp.149-160, (2001).
- [10] Ratnasamy, S., Francis, P., Handley, M., Karp, R., and Shenker, S. A scalable content-addressable network, In Proceedings of ACM SIGCOMM, Volume 31 Issue 4, Pages 161-172, (2001).

Reducing Probe Data in Telematics Services Using Space and Time Models

Ryozo Kiyohara[†], Hirohito Kakizawa[†],
Shinji Kitagami[‡], Yoshiaki Terashima[‡] and Masashi Saito[‡]

[†]Kanagawa Institute of Technology, Japan

[‡]Mitsubishi Electric Corp., Japan

{kiyohara@ic, s1021046@cce}.kanagawa-it.ac.jp
{mail52}@contact.kitagamishinji.net

{Terashima.Yoshiaki@eb, Saito.Masashi@bc}.MitsubishiElectric.co.jp

Abstract - In-vehicle information devices such as car navigation systems and smartphones are now widely used and they provide drivers with a lot of information, such as traffic jam information, weather forecast, etc., by a communication function such as cellular networks. These services gather a lot of information from cars or traffic sensors on the road, which results in many small pieces of data being transmitted over cellular networks. We propose a method that, by predicting car behavior, reduces the amount of such data. We observe the amount of traffic data, simulate vehicle behavior, and evaluate our models. Our conclusions show good results.

Keywords: Telematics service, ITS, Smartphone, Probe data, Car navigation System

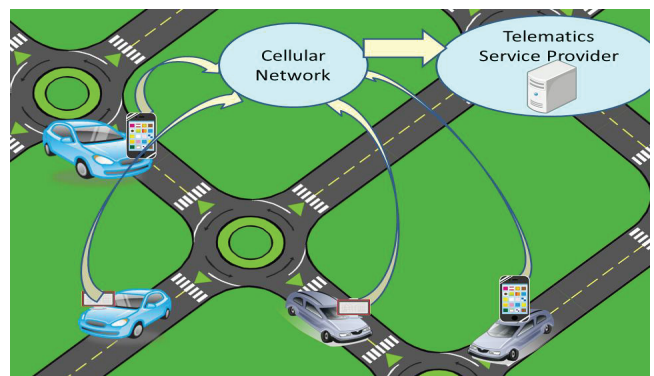


Figure 1: Telematics service

1 INTRODUCTION

Recent years have witnessed the emergence of many telematics services. For example, vehicle information devices can assess traffic conditions and display the fastest route to a destination. Such services provide information obtained from traffic sensors, historical traffic data, and in-vehicle information devices or smartphones, which are in widespread use [1][2][3][4].

Nevertheless, vehicle information devices for these services have to connect the cellular networks, and drivers have to pay additional costs for telematics services. If drivers use smartphones, then they do not have to pay these additional costs; however, the display sizes on these devices are too small. A new device called “Display Audio” [5] can display the same image with a smartphone connected by wireless or wired communication functions.

Therefore, smartphones can become popular telematics terminals that have several sensors and wireless communication functions[6]. Each smartphone frequently transmits a small amount of probe data. Therefore, a large amount of probe data can be transferred from vehicles to telematics service providers (TSPs). However, such increasing volumes of data traffic need to be regulated because of the high communication costs for both users and TSPs.

In a previous study [7], we proposed data compression methods for probe data by considering only the data itself, but we

were unable to reduce the amount of communication of control information that is a task that requires peak cutting methods. Therefore, in this paper, we propose a new method for reducing probe data on the basis of vehicle behavior prediction.

2 TELEMATICS SERVICE

2.1 What are Telematics Services

Telematics services are services that provide useful information to the driver. These services have to gather a lot of information from each in-vehicle device (See Figure 1). There are four types of telematics services:

1. A TSP gathers a large amount of information from in-vehicle information devices through the cellular network for each service; such information includes traffic jam information, weather information, route guidance, etc. The TSP then analyzes the data and delivers useful information to each in-vehicle information device.

2. If an accident takes place, the in-vehicle information device calls an emergency center, providing, automatically or manually, location information synchronized with the airbag information.
3. Entertainment services such as SNS, messaging, Internet, etc.
4. Vehicle relationship management (VRM), which gather a large amount of different types of information from the controller area network (CAN) for monitoring the vehicle status, etc.

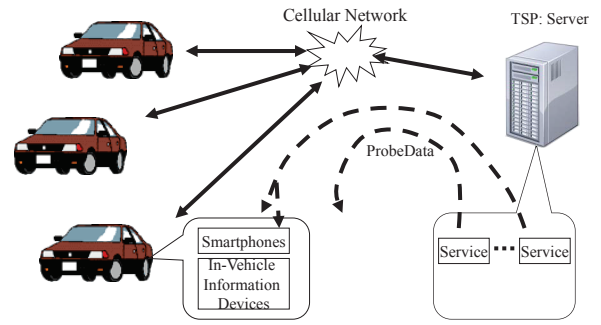


Figure 2: Probe data

Current in-vehicle information devices communicate with the TSP's server through a cellular network. Therefore, most of those services are able to run on the smartphones. Many drivers have smartphones and can use the cellular network at fixed costs and without having to pay additional money.

However, smartphones' displays are too small to show drivers a map or other information, which they cannot use when they are driving. In the near future, it is expected that display-audio devices, which show the smartphones' display image on the in-vehicle devices, will be used widely.

2.2 What are Probe Data

Probe data are gathered from many vehicles to the TSP's servers, as shown in Figure 2. There are three types of probe data:

1. Probe data that are gathered from various vehicles at short fixed intervals (e.g., 1 min and 5 min). This data includes average speed, travelling time, location information, and wiper information. The data are analyzed for traffic information, weather forecast, etc.
2. Probe data that are transported to the TSP or other services as soon as possible; these include airbag information and broken information for making emergency calls.
3. Probe data that should be stored and gathered to the manufactures for VRM; these include error logs in various Electronic Control units (ECUs), etc.

In this paper, we focus on the first type probe data. The size of each data is very small, but the number of communications is very large. Therefore, we have to solve the following three problems:

- Minimizing the probe data
- Minimizing the control information for each communication
- Minimizing the number of communications

In our previous research, we proposed a compression method for (1). In this paper, we propose a new method for reducing the number of communications in telematics services.

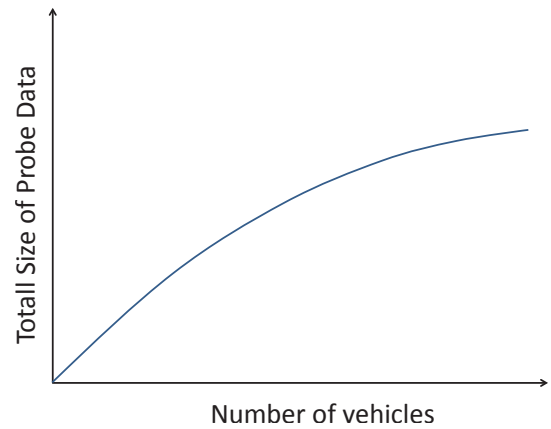


Figure 3: Number of vehicles and total size of probe data

2.3 Amount of Probe Data

Let x be the number of vehicles in a fixed area, which are able to communicate with the same base station through a cellular network. Let s be the size of probe data transmitted in a single communication. Let t be a fixed time. In this case, the total size T of the probe data is given by the following equation:

$$T = \sum stx \quad (1)$$

In our previous research [7], we proposed a data compression method. The data size depended on the number of vehicles. When the number of vehicles is small, the data are not compressed effectively; but when the number is large, the data can be compressed effectively. The relation between the number of vehicles and the total size of probe data is shown in Figure 3. If all in-vehicle information devices transmit probe data, the total size of probe data increases monotonically.

3 RELATED WORKS

There are many studies on the reduction of probe data in a cellular network. These studies may be categorized into three types:

1. Reducing the number of vehicle information devices that communicate to TSPs through a cellular network. In [9], inter-vehicle communication technologies (V2V) were used, and good results were obtained. However, few vehicles have the equipment required for V2V communications, which accounts for a significant problem. In [10], roadside communication (V2R) technology was used, but the covered area was very small.
2. Reducing the data size. Our previous method [7] showed good results, but it could not be used for peak cutting.
3. Controlling the number of vehicles driving in a fixed area. In [9], the TSP sent a message to a vehicle and guided it along its route. In this case, control of the vehicle was limited.

We thus propose a new space and time model for reducing the size of probe data without requiring V2V or V2R technologies.

4 METHOD OF REDUCING PROBE DATA SIZE

4.1 Proposed Architecture

Figure 4 shows the architecture of a telematics service system based on a telematics agent model that we propose. In this model, there are two types of agents: one runs on smartphones, whereas the other runs on the servers in a TSP.

The TSP agent monitors vehicular traffic in certain areas divided into zones. If the number of vehicles in a zone exceeds a certain threshold, the TSP agent selects smartphones according to the space-and-time strategy and instructs the smartphone agents to cease probe data transfer.

The basis of this idea is that the amount of probe data that is needed to provide a large amount of information is not very large. Therefore, we can select some appropriate vehicles. The number of selected vehicles is defined by two threshold values as follows (Figure 5).

1. If the number of smartphones is less than the threshold, all vehicles are selected.
2. If the number of smartphones is less than the threshold and greater than the threshold number1, a fixed number of vehicles are selected.
3. If the number of smartphones is greater than the threshold, a fixed numbers of vehicles are selected.

The number of selected vehicles is limited and fixed. Therefore, communication traffic through the cellular network is limited.

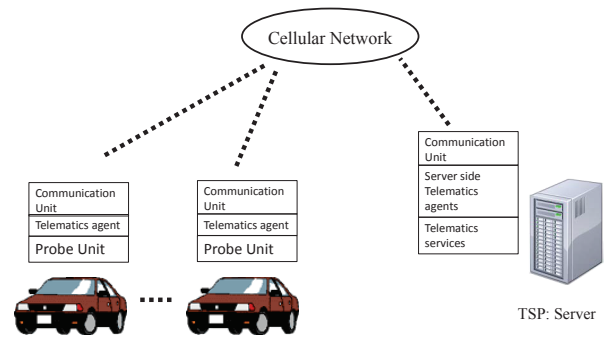


Figure 4: Proposed Architecture

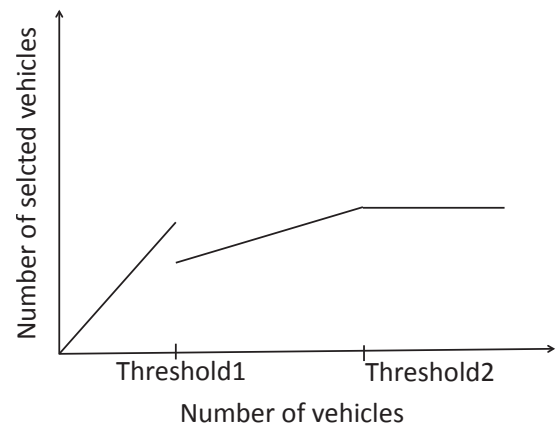


Figure 5: Number of vehicles and number of selected vehicles

4.2 Space and Time Strategy

The algorithm for selecting smartphones that should not transmit probe data is based on the following conditions.

1. In a particular area, a fixed number of smartphones are allowed to transmit probe data.
2. A smartphone transmits probe data only when its behavior cannot be predicted.

Thus, the volume of communication is limited. Actually, if there is less traffic in an area, then the number of smartphones will be relatively small. Hence, all the smartphones in the area can connect to the TSP. Conversely, if an area has heavy traffic, it will have a large number of smartphones. Therefore, only a fixed number of smartphones will be allowed to transmit probe data.

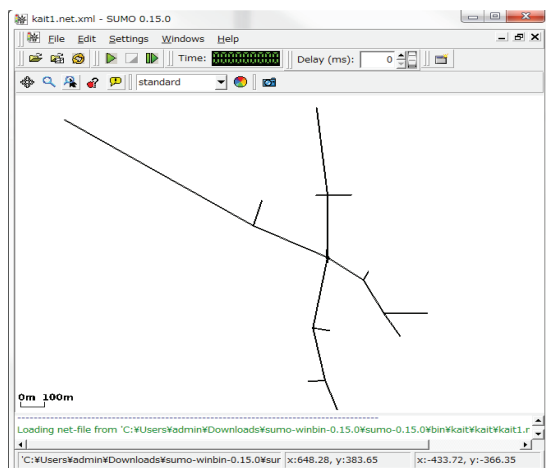


Figure 6: Atsugi City (near the Kanagawa Institute of Technology)

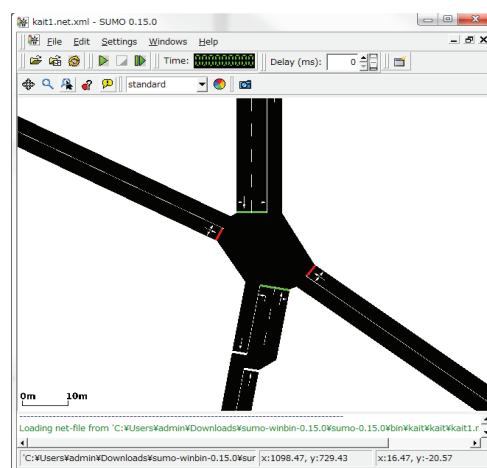


Figure 7: Atsugi City (near the Kanagawa Institute of Technology (zooming))

4.3 How is the Behavior of a Vehicle Predicted

The prediction algorithm for the behavior of a vehicle is as follows:

- If a smartphone uses a route guidance application, the TSP can predict the route. Therefore, the smartphone can communicate with the TSP only illegally.
- If the smartphone does not use the route guidance application, it is difficult to predict the route. However, in many cases, the vehicle may be on a long street. Therefore, the smartphone can communicate with the TSP either illegally or drive on any route except the main route.
- In a fixed area, the number of smartphones that communicate with the TSP is limited. Therefore, the TSP provides communication ratio as it decides.

5 EVALUATION

We evaluated our algorithm by carrying out a traffic simulation [11]. In a previous study [7], simulations were conducted using real traffic data and maps. We adopted the same approach in the present study to obtain accurate results.

5.1 Base Experimentation

Table 1 shows the average speed and Table 2 shows the average travelling time.

We observed the amount of traffic near the Kanagawa Institute of Technology. Figures 6 and 7s of that area.

On a weekday morning, the number of vehicles in an hour is 1,729, which is normal in that area and is not considered as crowding. The number of signal turns is 52.

In addition to these numbers, for the simulation, 2,593 (150%) vehicles constitute a crowded case, and 864 (50%) vehicles constitute a sparse case.

Table 1: Average Speed in that area

Sparse case	8.3 m/s
Nomarl case	7.5 m/s
Crowded case	5.3 m/s

Table 2: Average travelling time

Sparse case	196s
Nomarl case	216s
Crowded case	328s

We then changed the number of vehicles and obtained the probe data from each vehicle. Figure 8 travelling time in that area, and Figure 9 shows the average speed of each vehicle in that area.

5.2 Reducing the Probe Data Based on the Space Model

We get the probe data from selected vehicles by space model. Selected ratio in that area is from 3% to 100%. Table 3, 4, and 5 show the average speed and average travelling time from each simulation.

This result means we have to select the smartphone at least 20% in the crowded case.

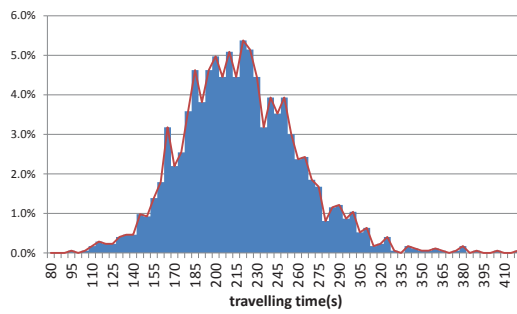


Figure 8: Atsugi City (Travelling time in that area)

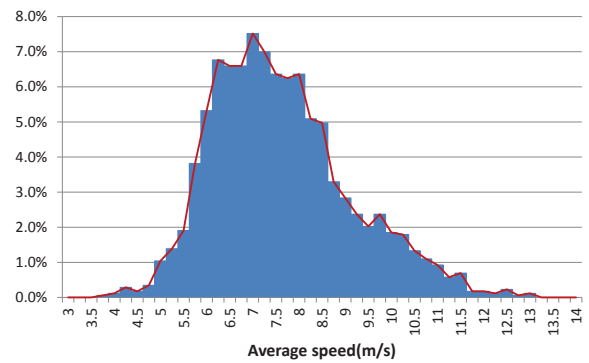


Figure 9: Average Speed in that area

5.3 Reducing the Probe Data Based on Time Model

We obtained probe data from the selected vehicles using the time model. The selection ratio for each vehicle varied from 20% to 100%. Tables 6, 7, and 8 show the average speed and average travelling time from each simulation.

The results indicate that we can reduce the number of selected smartphones in at least 80% of the normal cases in the time model. In other words, each smartphone should communicate with a TSP 20% of the time.

6 CONCLUSION

We proposed a new method for reducing the amount of probe data in telematics services. In addition, we confirmed the effectiveness of the proposed method by conducting simulations using real traffic data and maps.

In future work, the time and space models will be combined, and we will perform simulations on larger areas.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number 25330119

REFERENCES

- [1] G-Book: <http://www.prepaidmvno.com/capacity-carrier-db-2010-2/company-briefs/company-overview-toyota-g-book-japanese-mvno-service-by-toyota-motor-corporation/> <accessed on 1/6/2013>
- [2] CARWINGS: <http://www.nissanusa.com/innovations/carwings.article.html> <accessed on 1/6/2013>
- [3] Smartloop: <http://pioneer.jp/press-e/2007/0509-1.html> <accessed on 1/6/2013>
- [4] Onstar: <http://www.onstar.com> <accessed 1/6/2013>
- [5] Display audio : <http://www.toyota.ca/toyota/en/display-audio> <accessed on 1/6/2013>
- [6] Maekawa,T. Fujita ,A. Satou, and S. Kimura, " Usage of M2M Service Platform in ITS," NEC Technical Journal, Vol.6. No.4, pp.43-47(2011)
- [7] Nakase Y., Hiei T., Saito M., Kambe H., and Kiyohara R., " Reduction of the Amount of Probe -Data in Telematics Services," IEEE ICCE 2013, pp. 592-593(2013)
- [8] Chen B. and Cheng H. H., " A review of the applications of agent technology in traffic and transportation systems," IEEE Transactions on Intelligent Transportation Systems, Vol.11, No.2, pp.485-497 (2010)
- [9] T. Hung, H. Ikeda,K. Kuribayashi, and Nikolaos Voziatzis, " Reducing the Network Load in CREPEnvironment," Journal of Information Processing, Vol.19, pp.12-24(2011)
- [10] S. Adachi, R Ikeda, H. nishii, et al, " Compression Method for Probe Data," Proc. Of the 11th World Congress on ITS (2004)
- [11] Sumo: <http://sumo.sourceforge.net/> <accessed on 1/6/2013>

Table 3: Average speed and travelling time on space model (sparse)

Selecting ratio	Average speed	Average Travelling Time
100%	8.3m/s	195.8s
50%	8.3m/s	196.4s
20%	8.2m/s	197.4s
10%	8.4m/s	196.7s
5%	8.5m/s	194.1s
3%	8.7m/s	194.9s

Table 4: Average speed and travelling time on space model (normal)

Selecting ratio	Average speed	Average Travelling Time
100%	7.5m/s	216.5s
50%	7.5m/s	216.2s
20%	7.4m/s	220.3s
10%	7.5m/s	216.0s
5%	7.6m/s	216.8s
3%	7.1m/s	232.3s

Table 5: Average speed and travelling time on space model (crowded)

Selecting ratio	Average speed	Average Travelling Time
100%	5.3m/s	328.2s
50%	5.3m/s	326.2s
20%	5.4m/s	322.6s
10%	5.5m/s	318.1s
5%	5.6m/s	311.9s
3%	5.7m/s	315.1s

Table 6: Average speed and travelling time for each selecting ratio on time model (space)

100%	8.3m/s	195.8s
50%	8.3m/s	194.8s
33%	8.3m/s	194.0s
20%	8.3m/s	192.0s

Table 7: Average speed and travelling time for each selecting ratio on time model (normal)

100%	7.5m/s	216.5s
50%	7.5m/s	215.5s
33%	7.5m/s	214.5s
20%	7.5m/s	212.4s

Table 8: Average speed and travelling time for each selecting ratio on time model (crowded)

100%	5.3m/s	328.2s
50%	5.3m/s	327.2s
33%	5.3m/s	326.2s
20%	5.3m/s	324.0s

Reactive Load Balancing During Failure State in IP Fast Reroute Schemes

Kazuki Imura[†], and Takuya Yoshihiro[‡]

[†]Graduate School of Systems Engineering, Wakayama University, Japan

[‡]Faculty of Systems Engineering, Wakayama University, Japan
930 Sakaedani, Wakayama, 640-8510, Japan
{s131009, tac} @sys.wakayama-u.ac.jp

Abstract

To augment reliability of IP networks against link/node failure, several IP fast reroute schemes have been proposed so far. They proactively compute backup paths and activate them when failure occurs to prevent packets from losing at the failure link/node. However, it is known that the network performance considerably degrades in the failure state of IP fast reroute schemes, because congestion hot spots often appear near the failure link/node. In this paper, we propose a reactive load balancing method that can be applied to the major IP fast reroute schemes that covers single failure. Our scheme works when an IP fast reroute scheme is activating its backup paths, and reduces the degradation of network performance due to failure. In our load balancing scheme, with the overhead of a few bit field on packet header, we can largely reduce the performance degradation in the failure state and mostly keep the throughput as it was in the normal (no-failure) state.

1 Introduction

The Internet has grown as a social infrastructure in the world, for which high-level reliability is required. Even a short-time disruption of a network may reflect on significant cost because various indispensable communications depend on this high-speed network. However, when link or node failure occurs in an IP network, it is difficult to avoid service disruption for a certain time as long as we deploy a traditional routing protocols such as OSPF [1] and IS-IS [2]. Unfortunately, the frequency of failure is not low enough, as reference [3] reported.

To augment reliability of IP networks, several IP Fast Reroute (IPFRR) techniques have been proposed [4] [5] [6]. They proactively compute backup paths and activate them when failure occurs to prevent packet loss at the failure links or nodes. Typically, they cover every single link or node failure, so that path disruption in an IP network can be eliminated in every single link or node failure scenario. These IPFRR schemes complement the network performance during the failure state, i.e., during the time period until the failure is repaired.

However, with IPFRR schemes, it is known that using backup paths in face of failure brings congestion on links around failure, which degrades the performance of the network [7]. Although network resources may be insufficient in failure state, it is strongly desired to reduce the degradation level of the network performance.

In this paper, we propose a reactive load balancing method that works in the failure state of IPFRR schemes. Namely, the proposed method tries to reduce the degradation of network performance in the time period in which IPFRR is activating its backup paths. When congestion occurs around the failure, our method utilizes the unused backup paths of the IPFRR scheme to have a part of the traffic escape from the congestion. With this two-level rerouting, we make use of the unused resources of networks around failure to reduce the degradation of network performance. Through traffic simulation, we show that the network throughput in case of failure is considerably improved and it comes to be a comparable level to the normal state where failure is not present.

This paper is organized as follows. In Section 2, we describe several major IPFRR schemes that covers single failure, and present several existing approaches on the load balancing techniques over IP networks as well as IPFRR schemes. In Section 3, we describe the proposed method in detail, and give the result of traffic simulation in Section 4. Finally, we conclude the work in Section 5.

2 Load Balancing in IP Fast Reroute Schemes

2.1 IP Fast Rerouting and Their Modeling

To augment reliability of IP networks against failure, many IP fast reroute (IPFRR) schemes have been proposed. We first describe the literature of IPFRR schemes as an underlying technology of the proposed method.

IPFRR is a scheme that proactively computes backup paths to prevent packet loss due to failure. In IPFRR, if a router detects the failure of the next-hop link (or node), the backup paths are immediately activated to forwards packets, instead of forwarding them to the failed components. One of the well-agreed goals of IPFRR is to cover single link/node failure with low overhead. Several schemes that achieved single-failure coverage are proposed so far.

One of the major approaches for such IPFRR schemes is the *tunneling approach* such as NotVia [5] [8] [9]. NotVia computes a tunnel in advance for every single-failure scenario i.e., for protection against single-node failure, every node has its tunnel that reaches the next-next-hop node on the shortest path without going through the next-hop (i.e., failure) node. When a packet meets failure, the packet is encapsulated to be forwarded into the tunnel to bypass the failed component.

Another major approach is the *two-table approach*, in which the secondary routing table is proactively computed to be ac-

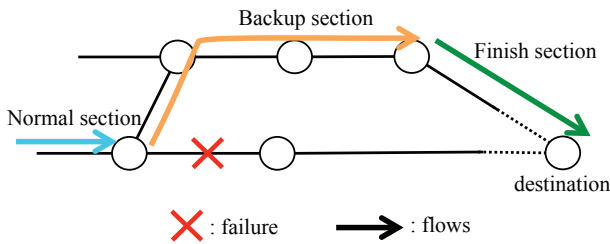


Figure 1: Model of IP Fast Reroute Schemes

tivated in case of failure. FIFR (Failure Inferencing based Fast Rerouting) [4] [10] would be the representative proposal in this approach, which covers every single link/node failure scenario. In FIFR, routers infer failure according to the in-coming network interface of packets, i.e., if a packet for a destination arrives from an unusual interface, the packet is forwarded using the secondary routing table to avoid the failure node. To improve the computational cost of FIFR, Xi, et al. proposed a time-efficient algorithm to compute the secondary routing table [11].

As an extension of the two-table approach, SBR (Single Backup-table Rerouting) was proposed [12] [13] [6]. SBR, which covers single link/node failure scenarios, achieved to prevent packet loops in case of multiple-failure to improve reliability of networks by utilizing a 2-bit field on packet header.

In this paper, we treat most of those IPFRR schemes that covers single link/node failure as an integrated model. In this model, (i) we first assume that every node uses a shortest-path based algorithm to compute the primary routing table, and (ii) every node has a backup path to forward packets without visiting the next-hop link or node. Also, (iii) the traveling paths of packets in these schemes can be splitted into three sections as shown in Fig. 1. I.e., the first section is the path along the shortest path that is used by the packets that have not met failure, the second section is the path along the backup routing configuration that is used after packets meet failure, and the third section is the path along the shortest path that the packets escaped from the second section uses to reach their destination. We call those three sections as *normal*, *backup*, and *finish* sections, respectively.

we hereafter treat this integrated model rather than treating each of IPFRR schemes in order to propose a general framework of load balancing for IPFRR.

2.2 Related Work on Load Balancing in IP Networks

Several load balancing methods have been proposed for shortest-path based IP networks, which is independent of IP fast reroute schemes. For instance, Antić, et al. proposed TPR (Two Phase Routing) [14], which distributes traffic into several paths using intermediate nodes, i.e., TPR once forwards packets to some intermediate nodes using IP tunnels and then forwards them to their destinations using normal shortest paths. Mishra et al. proposed S-OSPF (Smart-OSPF) [15], in which source nodes distribute traffic to their neighbor nodes to balance traffic load among those neighbors, i.e., among the shortest paths that start from the neighbors to a destination.

They work effectively with relatively low overhead. However, in case of failure under IPFRR schemes, they do not work well immediately because they have to decide the distribution ratio of traffic among several possible paths based on the measured traffic load given as the demand matrix of the network. Namely, when failure occurs, they first have to measure the traffic load over the network, and then compute the optimal distribution ratio. This process inevitably includes considerable delay before these load balancing schemes work effectively, even though IPFRR schemes achieves 50msec recovery from failure.

As a load balancing scheme using backup paths of the IPFRR schemes, Hara et al. proposed a method to utilize backup paths of IPFRR not only the failure state, but also in the usual (no-failure) state to provide load balancing functionality [16]. In their scheme, they watch congestion using the output queue length of routers, and when congestion is detected in the primary link, packets are rerouted into the backup path. Note that the rerouted packets may cause other congestion. If the congestion invokes another rerouting of packets, this endless chain of rerouting leads harmful confusion over the network. To prevent this, the rerouted packets are given less priority than the original packets, and they are dropped by priority in face of congestion. This method can utilize backup paths effectively even in case of the normal state. However, their load balancing functionality does not work in the failure state because backup paths are occupied to protect routes against failure.

There is a few load balancing method that works in the failure state of IPFRR schemes. For instance, Ho, et al. proposed a new IPFRR scheme with load balancing functionality in which packets are distributed to several intermediate nodes using tunnels when a flow meets failure [17]. In [18], Zhang, et al. also proposed a post-failure load balancing method over LFA (Loop-Free Alternate) [19], in which packets are distributed to several neighbor nodes. They compute the distribution ratio over several backup paths based on the measured traffic matrix. Thus, they require the overhead and the delay to estimate the traffic matrix. Furthermore, their IPFRR schemes do not provide full coverage of single failure.

In the current state of the art, no load balancing method seems to be present that works in the failure state of full-coverage IPFRR schemes. We in this paper propose the first one that distributes traffic load by reusing backup paths of the major full-coverage IPFRR schemes.

2.3 Base Load Balancing Scheme over IPFRR Schemes

In this section, we give an in-depth description of the load balancing scheme of Hara et al. [16]. The method that we propose in this paper is an extension of their method.

As mentioned in Sec. 2.1, packets in IPFRR schemes travel through the three sections of the forwarding paths. In [16], they utilize a 2-bit field in packet header to mark packets to know which section the packets are travelling. We call this mark the *state* of packets, and each state corresponding to the three sections of paths is called as *normal*, *backup*, and *finish* states.

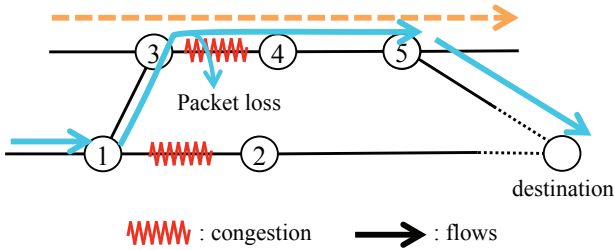


Figure 2: Load Balancing using IP Fast Reroute [16]

For the mechanism of [16], see Fig. 2. When a packet of *normal* state meets congestion (at router 1), it is forwarded into the backup section with its state changed to *backup* state. To detect congestion (at router 1), the router watches the length of the output queue: if the queue length for the primary next-hop interface is longer than threshold T_1 , the router detects the congestion, and enqueue the packet into the queue of the backup interface. (See Fig. 3(a) for the behavior of router 1.) When the packet of *backup* state reaches the end of the backup section (at router 5 in Fig. 2), the state is changed to *finish* state to be forwarded along the shortest path to reach their destination.

As mentioned in Sec. 2.2, the packets forwarded into the backup section may cause other congestion. To prevent the harmful influence of the rerouting chain, this scheme gives less priority to the *backup*-state packets to have those packets dropped as soon as the packets meet congestion again. For the specific mechanism, see Fig. 3(b). When the *backup*-state packets reach Router 3, if the queue length for the backup-path interface is longer than threshold $T_2 (< T_1)$, the *backup*-state packets that are given less priority are silently dropped. It is shown in [16] that the mechanism scarcely reduces the performance of the original flows in both delay and throughput even in a high-load network.

The *finish* section is almost the same as *normal* section. The priority of *finish*-state packets is as usual so that T_2 threshold is not applied to them. However, if we allow reroutes for *finish*-state packets, it may cause routing loops in case of multiple failures. Thus, we can prohibit rerouting of *finish*-state packets, or, as is proposed in [16], limit the number of rerouting that a packet can experience.

3 Load Balancing Method for Failure State

3.1 Basic Idea

We propose a reactive load balancing method for IPFRR that works in the failure state by extending the mechanism of [16]. Our basic idea is to allow secondary rerouting of packets to reduce congestion caused by the rerouting packets coming from failure.

The idea of this two-level rerouting is illustrated in Fig. 4. Assume that failure of link (1, 2) occurs and packets are forwarded into the backup section at node 1. Also assume that the rerouted packets arrive at node 3 and a link (3, 4) is congested as a result. With the conventional method [16], the rerouted packets are likely to be dropped because the rerouted packets are given less priority. This leads the result where

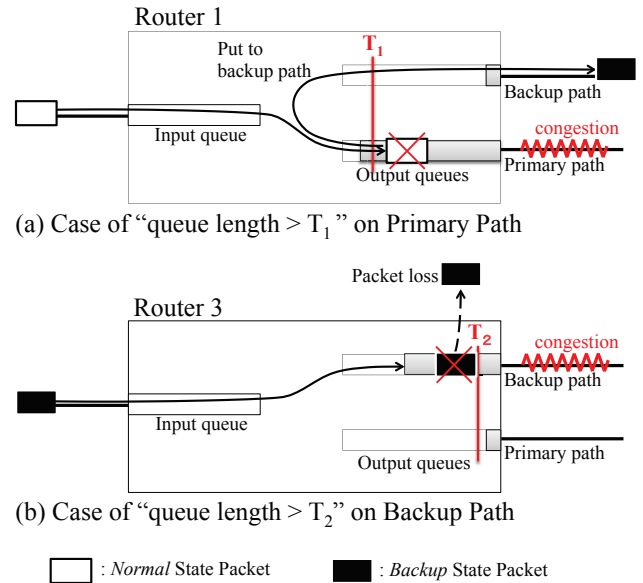


Figure 3: Load Balancing Behavior Based on Queue Length

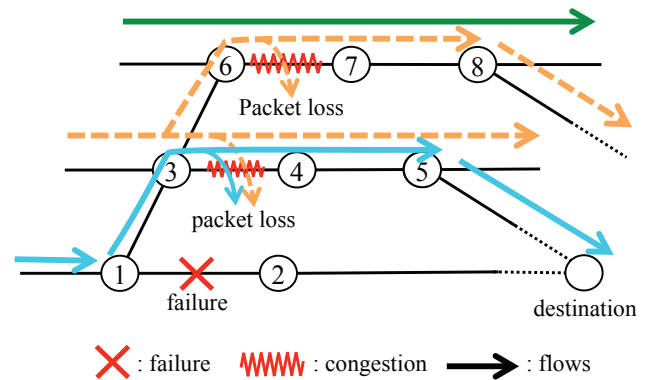


Figure 4: Two-level Rerouting of the Proposed Method

only a part of rerouted packets can be saved by IPFRR, which degrades the network performance.

In contrast, our new method allows to reroute the original traffic on node 3 that has been using the link (3, 4). This secondary rerouting makes room on link (3, 4) for the rerouted packets that come from router 1, which shrinks the congestion on link (3, 4). In other words, we utilize the resources of the secondary backup path to reduce the degradation of network throughput that caused by the congestion on link (3, 4). Note that we do not allow the third rerouting to prevent the harmful chain of congestion. This is because the rerouting of packets includes both benefits and risks over the performance in communications; If we allow higher-level rerouting, then the throughput may be improved by utilizing the backup paths of larger numbers of nodes, while the redundant rerouted traffic that degrades communication quality may increase over the wider area of the network. Because we regard it important to clarify the benefit of the proposed method with the most elementary condition, we in this paper examine the two-level case first. Specifically, we give less priority to the secondary rerouted traffic in the same way as [16] to have them immediately dropped on the congested link (6, 7).

3.2 State Transitions

In the conventional scheme [16], they use three states (i.e., *normal*, *backup*, and *finish*) of packets to perform priority processing of packets. In our scheme, because we have to distinguish the primary rerouted packets from the secondary ones, we introduce another state: when a packet is rerouted due to failure (i.e., the first-level rerouting is invoked at router 1 in Fig. 4), the packet state changes to *backup* state. And, when a packet of *normal* state is rerouted due to the followed congestion, (i.e., the second-level rerouting invoked at router 3 in Fig. 4), the packet state changes to *secondary* state.

Note that we in this paper do not distinguish the backup section and the finish section of the rerouting path; the both sections in the first-level reroute correspond to the *backup* state, and those in the second-level correspond to the *secondary* state. This is because, as described in Sec. 2.3, the role of the *finish* section in [16] is to limit the number of rerouting that a packet experience. This function plays an important role when we deploy a load balancing function in the practical scenes, however, to simplify the description of our scheme in this paper, we do not refer to this function hereafter. I.e., we regard that the *backup* section and the *finish* section correspond to the same state.

3.3 Packet Priority Control

In our scheme we give each packet a proper priority to prevent the harmful congestion chains as well as to take the fairness among flows into account.

We first point out that, at the congestion link that caused by failure (e.g., link (3, 4) in Fig. 4), the original flows and the rerouted flows should be treated equally. It is because failure should be concealed from users in IPFRR schemes, and consequently those flows should be fairly processed. Thus, we do not apply the rule of T_2 threshold (described in Sec.2.3 and Fig. 3(b)) to both the *normal*- and *backup*- state packets. It means that the drop probability of *normal* and *backup* packets is the same in the congested link.

On the contrary, we have to prevent the rerouting chain that causes harmful congestion so that we limit rerouting within two levels. Specifically, at the congested link that is caused by the first-level rerouting (e.g., link (6, 7) in Fig. 4), we apply the rule of T_2 threshold into *secondary*-state packets to drop them immediately when they meet another congestion.

The formal packet forwarding rule is shown in Table 1. This table shows the behavior of routers when a packet comes in, under every possible conditions.

3.4 Enabling Secondary Rerouting

In our scheme, we enable the secondary rerouting only when failure occurs. In the practical scenes of network management, the multi-path load balancing mechanism includes several inconveniences such as large jitter or packet reordering. Thus, many network operators would not wish to use the load-balancing function as usual. To reduce such degradation of networks in normal state, we introduce a mechanism to enable the secondary rerouting mechanism only when the first-level rerouting is occurring.

Table 1: Forwarding Rule of Proposed Scheme

packet state	Conditions		Operations	
	failure	queue length	forward to	state change
<i>normal</i>	No	$l(p) \geq T_1$	backup Path	To <i>secondary</i>
		$T_1 > l(p)$	primary path	-
	Yes	-	backup path	To <i>backup</i>
<i>backup</i>	No	$l(b) \geq T_1$	drop	-
		$T_1 > l(b)$	backup path	-
	Yes	-	drop	-
<i>secondary</i>	No	$l(b) \geq T_2$	drop	-
		$T_2 > l(b)$	backup path	-
	Yes	-	drop	-

* $l(p)$: queue length for primary path, $l(b)$: queue length for backup path

* "Failure" means the failure existence on primary nexthop for *normal*-state packets, on backup nexthop for *backup*- or *secondary*-state packets.

Table 2: Simulation Settings in Random Network

Item	Value
Topology	Waxman Model
#Nodes	30
#Links	60
Link Bandwidth	1.0 Mbps
Link Delay	1 msec
Flow Type	CBR
Rate of a Flow	200 kbps
Packet size	1 kbytes
Output Queue size	50 packets
Threshold T_1	10% of queue size
Threshold T_2	100% of queue size
#flows	20-150

To control the load-balancing function, we introduce a 1-bit flag for each interface on every router that indicates the load-balancing is enabled or not. I.e., only if the flag is *true*, a *normal*-state packet that uses the corresponding interface as its primary next-hop interface can be forwarded into the backup section with its state changed to *secondary* state, when it meets congestion.

The 1-bit flag is initially *false*, and is changed to *true* when a *backup*-state packet that uses the corresponding interface arrives at the router. The flag becomes *false* when no such *backup*-state packet arrives at the router in the past certain time.

Table 3: Simulation Settings in Rocketfuel Network

Item	Value
ISP Name	Telstra EBORN Tiscali Exodus Abovenet
#Nodes	104 87 161 79 138
#Links	151 161 328 147 372
LinkBandwidth	5 / ilink weightj Mbps
Link Delay	Obtained from Rocketfuel
Flow Type	CBR
Rate of a Flow	200 kbps
Packet size	1 kbytes
Output Queue size	50 packets
Threshold T_1	10% of queue size
Threshold T_2	100% of queue size
#flows	50-150

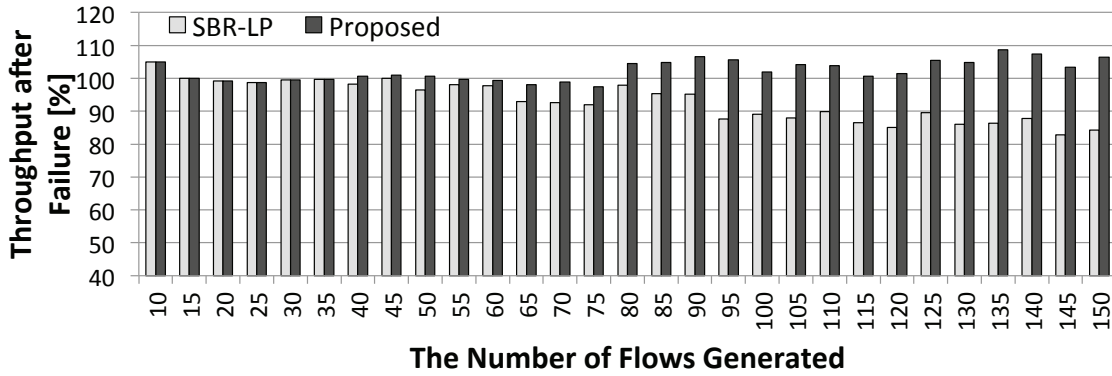


Figure 5: Throughput with Various Random Network Load (Average of 30 trials)

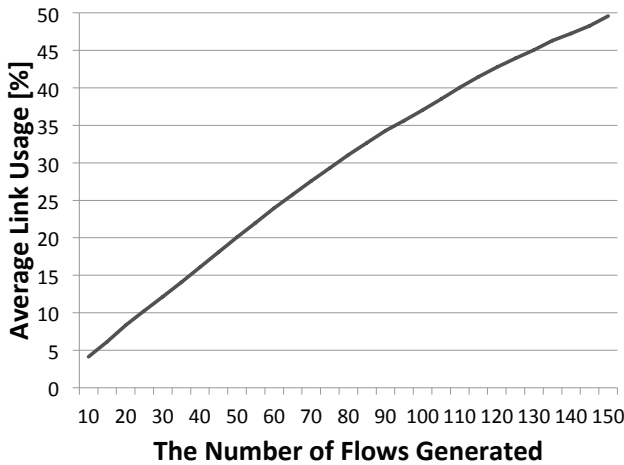


Figure 6: Average Link Occupancy in Normal State

4 Performance Evaluation

4.1 Simulation Scenario

We evaluate the proposed method using the ns-2 simulator [20]. As a base IPFRR scheme, we chose a link protection scheme SBR-LP (Single Backup-table Rerouting - Link Protection) [13] because SBR-LP is convenient to cooperate with the proposed load balancing method in that it utilize a 2-bit field of packet header. Note that SBR-LP provides backup paths against every single link failure, in which the algorithm to compute backup paths is similar to [11]. We implemented SBR-LP and the proposed method over SBR-LP on ns-2.

In this evaluation, we designed the scenarios in which two different types of topologies are used. In scenario 1, we use the randomly generated topologies that model the Internet. In scenario 2, we use the topologies of real networks retrieved from a public database.

First, we describe scenario 1. The underlying network is a random network based on Waxman [21] model topology generated by a topology generator BRITe [22], where the number of nodes is 30 and that of links is 60. The bandwidth of every link is 1.0 Mbps and its transmission delay is 1 msec. Note that, although the link speed is extremely low compared to the practical networks, we can estimate the throughput of high-speed networks using the proportion of link speed. We generate CBR flows randomly, i.e., we select a source and

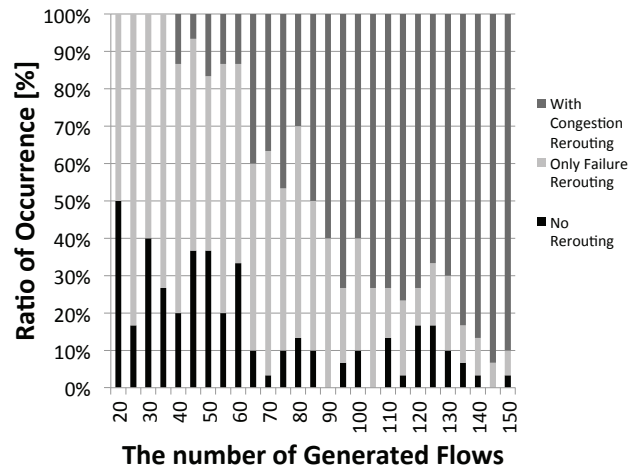


Figure 7: Ratio of Rerouted Flows in Proposed Scheme

destination node randomly from the network. The packet size is 1 kbytes and the transmission rate of every flow is 200 kbps. The output queue size is as large as 50 packets. Threshold T_1 and T_2 is set at 10% and 100% of the queue capacity each. We vary the number of flows generated between 20 and 150 with interval of 5. For each case of flow numbers, we performed 30 trials of the simulation; we use 5 different random topologies with 6 different random seeds. The overview of simulation settings in scenario 1 are shown in Table 2.

On the other hand, in scenario 2, evaluation is done through simulations using real network topologies. We used five topologies, Telstra, EBONE, Tiscali, Exodus, and Abovenet, obtained from the topology database site Rocketfuel [23]. We set the bandwidth of every link to be in inverse proportion to its weight. Specifically, the link bandwidth is set to $5/\text{weight}$ Mbps. Flows are generated in the configuration similar to S-scenario 1. The transmission rate of every flow is 200 kbps. The packet size is 1 kbytes and the output queue size is as large as 50 packets. Threshold T_1 and T_2 are set at 10% and 100% of the queue capacity, respectively. We vary the number of flows generated between 50 and 150 with interval of 10. For each case of flow numbers, we performed 12 trials of the simulation with different random seeds. The overview of the simulation settings in scenario 1 is shown in Table 3.

In both scenarios 1 and 2, we measure the performance in case of single link failure, i.e., the failure scenario that IPFR-

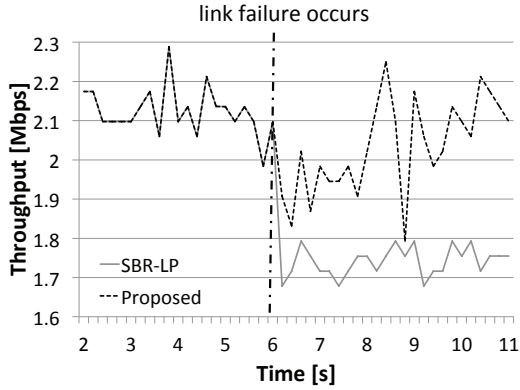


Figure 8: Throughput in Time Series

R schemes can cover. To this end, we generated flows at 1 second from start, and make a single link failure randomly 5 seconds later. We measured the throughput of 4-second periods of both before and after the failure occurs; we measured the time period between 2-6 seconds from start as the performance of the normal state, and measured the time period between 7-11 seconds from start as that of the failure state. Note that, we did not measure the throughput of all flows, but we measured the throughput of the flows that are affected by the failure. Specifically, we defined the *affected flows* as the flows whose packets are rerouted by the proposed scheme in the failure state, and measured the throughput of them to compare the performance of two schemes.

4.2 Results

4.2.1 Random Topology Scenarios

We describe the result of scenario 1. Fig. 5 shows the average throughput of the *affected flows* for each case of flow numbers in the failure state. The throughput is shown as the ratio compared to the throughput in the normal state, i.e., 100% means the same throughput as the normal state. Although the throughput of the conventional method (i.e., SBR-LP) decreases as the flow number increases, the throughput of the proposed method keeps the same level. Note that in several cases the throughput of the proposed method in the failure state is larger than the normal state, i.e., it exceeds 100% in Fig. 5. This is because the congestion that was occurring in the normal state can be reduced by the proposed method, due to the mechanism described in Sec. 3.4, which enables the secondary rerouting only when failure occurs.

Fig. 6 indicates the average link occupancy of each case of flow numbers in the normal state, and Fig. 7 shows the ratio of rerouted flows, i.e., the ratio of the flows that did not experience reroutes, that experienced reroutes caused by failure, and that experienced reroutes caused by congestion, in the failure state of the proposed scheme. These figures show that, as the link occupancy grows, the flows rerouted due to congestion increase, and accordingly the difference of throughput between the conventional and the proposed schemes grows larger.

Fig. 8 shows the time series of throughput measured with 0.2 second interval, in a typical case of 150-flow scenario. As

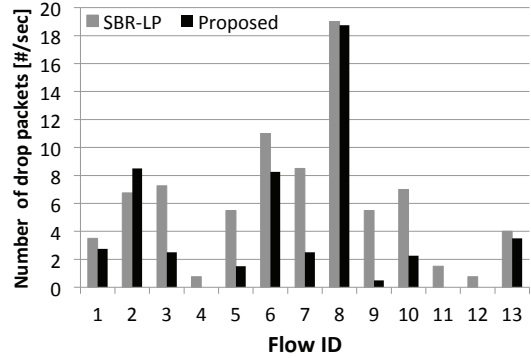


Figure 9: Ratio of Drop Packets per Flow

the figure shows, the throughput of SBR degrades immediately when failure occurs, while the proposed method keeps the same level of throughput. It shows that the proposed scheme works immediately after failure, without delay. Note that the throughput of the proposed method in a few second just after failure is a little lower than following several seconds. This is caused by the queueing delay that rapidly increases when failure occurs due to the congestion on the backup paths. Although the effect of the queueing delay is kept in a few second in this scenario, the queueing delay in practice will shrink significantly as the link speed is far larger than that.

Fig. 9 shows the ratio of drop packets of the *affected flows* in the same scenario. Although there are several exceptions, most flows reduce the drop packet ratio compared to SBR. This is the gain brought from the secondary rerouting.

4.2.2 Real Topology Scenarios

We describe the result of scenario 2. Figs. 10-14 show the throughput of the *affected flows* in the failure state in each of five topologies. Same as Fig. 5, these figures show the ratio of the throughput in the failure state compared to the normal state

Throughput of the proposed method is higher than SBR-LP in all topologies. These results show that the proposed method is also effective in the real topology.

In both SBR-LP and the proposed method, there are the cases where the throughput after failure exceeds the throughput before failure. In SBR-LP, such a case occurs when the rerouting packets reduce the traffic in the shortest paths, and consequently reduce congestion in these paths. In the proposed method, this tendency goes greater than SBR-LP. This is because the secondary rerouting in the proposed method often reduces the congestion that have been occurring even in the absence of link failure.

4.3 Discussion

We proposed a load balancing method for the failure state of IPFRR schemes. In the state where packets are rerouted by IPFRR against failure, the proposed method further reroutes packets at the congestion that are generated by the rerouted packet, using detour paths of IPFRR to reduce the conges-

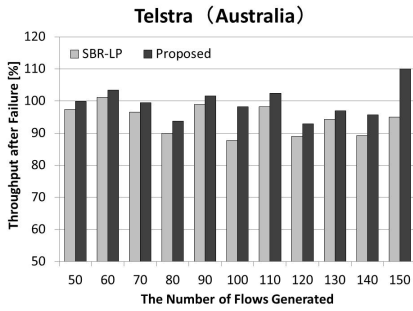


Figure 10: Throughput with Various Telstra Network Load (Average of 12 trials)

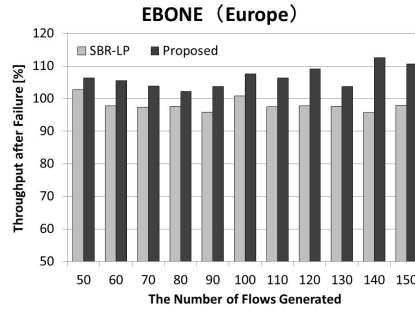


Figure 11: Throughput with Various Tiscali Network Load (Average of 12 trials)

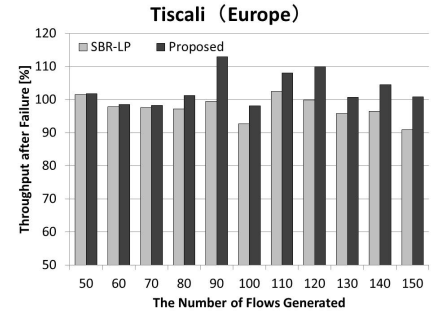


Figure 12: Throughput with Various Exodus Network Load (Average of 12 trials)

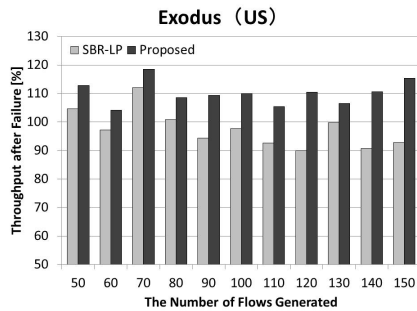


Figure 13: Throughput with Various Abovenet Network Load (Average of 12 trials)

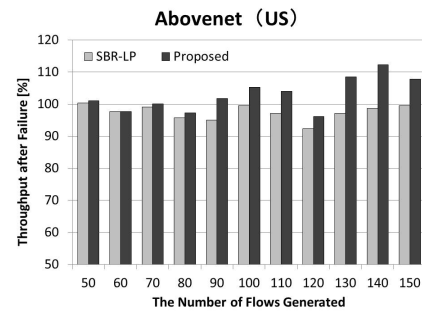


Figure 14: Throughput with Various Abovenet Network Load (Average of 12 trials)

tion. Through the evaluation, we confirmed that the proposed method reduces the congestion that occurs in the backup paths, and prevents the degradation of throughput performance in case of single failure.

Note that, since the proposed method reduces the congestion that occurs on backup paths using secondary rerouting, the proposed method compels flow, that are not directly affected by link failure, to reroute. Thus, the proposed method causes larger jitter and packet reordering by a part of packets in these flows rerouted.

Through the simulations, we have shown that the throughput in the failure state is as large as, or more than that of the normal state with the proposed method. Throughput is actually the most important criterion to measure network performance and so the proposed method is effective to improve the network performance. However, the degradation of communication quality from the viewpoint of jitter and packet reordering is not still negligible.

For this kind of quality degradation, several methods that reduce the harmful influence of packet reordering have been proposed in the literature. In general, in this area of research, the granularity of load distribution has been discussed as one of the essential tradeoff issues; if we distribute traffic in the per-packet fashion, the effect of reordering goes significant, and if we do it with larger granularity, the load balancing performance degrades. The traditional method to reduce the effects of reordering is to use the granularity of flows instead of that of packets, using a hash function [24] [25]. Several studies provide more sophisticated packet manipulations

for traffic distribution that are robust against packet reordering [26] [27]. New TCP protocols to endure packet reordering are also presented [28]. One of the future tasks is to apply them to the proposed method, and evaluate the performance against packet reordering.

5 Concluding Remarks

In this paper, we proposed a reactive load balancing scheme to reduce degradation of network performance in the failure state of IPFRR schemes. Our scheme works over major IPFRR schemes that covers single failure, and reuses their backup paths reactive to distribute traffic into wider area of networks around the failure component. Therefore, it requires small overhead of packet marking compared to the existing load balancing techniques, which instead requires the overhead of estimating traffic matrix. The evaluation using a network simulator clarified that we can achieve at least the same level of the throughput as the normal state even in the failure state.

One of the problems in this scheme would be the problem of packet reordering. If a traditional TCP is used in this network, the packet reordering degrades the throughput of flows, which reduce the effect of the proposed scheme. Accordingly, one of the future tasks is to apply existing techniques [26] [27] [28] to reduce the harmful influence of packet reordering, and to evaluate the performance of the proposed method against packet reordering. To grasp the level of harmful influence of packet reordering in comparison with the good aspect of the proposed method over throughput is one of the essential task

for the future.

REFERENCES

- [1] J. Moy, "OSPF Version 2," IETF RFC2328, April, 1998.
- [2] ISO/IEC, "Intermediate System to Intermediate System intra-domain routing information exchange protocol for use in conjunction with the connectionless-mode network service (ISO 8473)," ISO/IEC, Tech. Rep. 10589:2002(E), April 2002.
- [3] A. Markopoulou, G. Iannaccone, S. Bhattacharyya, C.N. Chuar, and C. Diot, "Characterization of failures in an IP backbone network," *IEEE/ACM Trans. on Netw.*, Vol.16, Issue 4, pp.749-762, 2008.
- [4] Z. Zhong, S. Nelakuditi, Y. Yu, S. Lee, J. Wang, and C.N. Chuah, "Failure inferencing based fast rerouting for handling transient link and node failures," in *Proceedings of IEEE Global Internet*, Mar. 2005.
- [5] M. Shand, S. Bryand and S. Previdi, "IP Fast Reroute Using Not-via Addresses," draft-ietf-rtgwg-ipfrr-notvia-addresses-04.txt, 2009.
- [6] T. Yoshihiro and M. Jibiki, "Single Node Protection without Bouncing in IP Networks," *IEEE 13th Conference on High Performance Switching and Routing (HPSR2012)*, pp.88-95, 2012.
- [7] S. Dasgupta, J.C.de Oliveira, and J.P. Vasseur, "A Performance Study of IP and MPLS Traffic Engineering Techniques under Traffic Variations," *IEEE Globecom2007*, pp.2757-2762, 2007.
- [8] A. Li, P. Francois and X. Yang, "On Improving the Efficiency and Manageability of NotVia," In *proc. of ACM CoNext 2007*, 2007.
- [9] G.Enyedi, P.Szilágyi, G.Rétvári, and A.Császár, "IP Fast ReRoute: Lightweight Not-Via without Additional Addresses," In *proc. of IEEE INFOCOM2009*, pp.2771-2775, 2009.
- [10] J. Wang and S. Nelakuditi, "IP Fast Reroute with Failure Inferencing," In *proceedings of SIGCOMM workshop (INM) 2007*, pp.268-273, 2007.
- [11] K. Xi and H.J. Chao, "IP Fast Rerouting for Single-Link/Node Failure Recovery," In *Proc. of IEEE BROADNETS2007*, pp.142-151, 2007.
- [12] H. Ito, K. Iwama, Y. Okabe, T. Yoshihiro, "Single backup table schemes for shortest-path routing," *Theoretical Computer Science*, 333(3):347-353, 2005.
- [13] T. Yoshihiro, "A Single Backup-Table Rerouting Scheme for Fast Failure Protection in OSPF," *IEICE Transactions on Communications*, Vol. E91-B, No. 9, pp.2838-2847, 2008.
- [14] M. Antić and A. Smiljanić, "Oblivious Routing Scheme Using Load Balancing Over Shortest Paths," In *Proc. IEEE ICC 2008*, pp.5783-5737, 2008.
- [15] A.K. Mishra and A. Sahoo, "S-OSPF: A Traffic Engineering Solution for OSPF based Best Effort Networks," In *Proc. IEEE Globecom 2007*, pp.1845-1849, 2007.
- [16] M. Hara and T. Yoshihiro, "Adaptive Load Balancing based on IP Fast Reroute to Avoid Congestion Hot-spots," *IEEE International Conference on Communications (ICC 2011)*, pp.1-5, 2011.
- [17] K. Ho, N. Wang, G. Pavlou, C. Botsiaris, "Optimizing post-failure network performance for IP Fast ReRoute using tunnels," *5th International ICST Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine08)*, 2008.
- [18] M. Zhang and B. Liu, "Traffic Engineering for Proactive Failure Recovery of IP Networks," Volume 16, Issue 1, Pages 5561, 2011.
- [19] A. Atlas and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-free Alternate," IETF RFC 5286, 2008.
- [20] "The network Simulator NS-2," <http://www.isi.edu/nsnam/ns/>.
- [21] B. Waxman, "Routing of Multipoint Connections," *IEEE J. Select. Areas Commun.*, December 1988.
- [22] A. Medina, A. Lakhina, I. Matta, and J. Byers, "BRITE: An approach to universal topology generation," in *Proceedings of IEEE MASCOTS*, pp.346-353, Aug, 2001.
- [23] N. Spring, R. Mahajan, D. Wetherall, and T. Anderson, "Measuring ISP topologies with rocketfuel," In *ACM SIGCOMM'02*, pp. 133145, Aug. 2002.
- [24] D. Thaler and C. Hopps, "Multipath Issues in Unicast and Multicast Next-Hop Selection," IETF RFC2991, 2000.
- [25] C. Hopps, "Analysis of an Equal-Cost Multi-Path Algorithm," IETF RFC2992, 2000.
- [26] S. Kandula, D. Katabi, S. Sinha, and A. Berger, "Dynamic Load Balancing Without Packet Reordering," *ACM SIGCOMM Computer Communication Review*, Vol.37, Issue 2, pp. 51-62, 2007.
- [27] R. Martin, M. Menth, and M. Hemmkeppler, "Accuracy and Dynamics of Hash-Based Load Balancing Algorithms for Multipath Internet Routing," In *Proc. IEEE International Conference on Broadband Communication, Networks, and Systems (BROADNETS)*, 2006.
- [28] S. Bohacek, J.P. Hespanha, J. Lee, C. Lim, and K. Obraczka, "A new TCP for persistent packet reordering," *IEEE/ACM Transactions on Networking (TON)*, Vol.14, Issue 2, pp.369-382, 2006.

A DTN Routing Scheme Based on Publish/Subscribe Model

Ryosuke Abe^{*}, Yoshitaka Nakamura^{**}, and Osamu Takahashi^{**}

^{*}Graduate School of Systems Information Science, Future University Hakodate, Japan

^{**}School of Systems Information Science, Future University Hakodate, Japan
{g2112002, y-nakamr, osamu}@fun.ac.jp

Abstract- Delay Tolerant Networking (DTN) is attractive as an effective communication method in unstable network environments where frequent disconnections occur easily. DTN routing is based on the store-carry-forward paradigm. So far, various DTN routing schemes corresponding to the temporal and spatial characteristics of contacts between nodes have been proposed. However, name resolution between a source and a destination is difficult in a network environment that consists of only wireless terminals such as DTN. In this paper, we present a DTN routing scheme based on the publish/subscribe model that enables flexible communication by using topics of information. In the proposed scheme, messages are sorted by the subscription lists and the contact condition of nodes in order to deliver to destinations with a short delay. We compare the performance of the proposed scheme with that of existing schemes through simulations.

Keywords: Delay Tolerant Networking, Store-Carry-Forward, Publish/Subscribe Model.

1 INTRODUCTION

Recently, as a development of near field communication technology and mobile devices, network services are becoming available in areas where a communication infrastructure is not set up and disaster areas. However, if nodes move frequently in such environments, frequent disconnections occur easily, so users cannot use the networks continually.

Delay Tolerant Networking (DTN) is attractive as an effective communication method in such unstable network environments [1]. DTN is intended to optimize communication performance and share network resources. To reach these goals, source nodes, relay nodes, and destination nodes work together and control the transmission of information. DTN has been primarily studied as a technology to be applied to communication in the sea, space, and disaster areas etc. However, in recent years, the number of applications and experiments that use DTN technology is increasing, such as communication in developing countries and the delivery of local news and advertising.

DTN routing is based on the Store-Carry-Forward paradigm [2]. In this paradigm, each node moves while keeping messages until it becomes possible to communicate with other nodes. When it meets the other nodes, it forwards replications of the message to them.

Generally, in DTN routing schemes, the message delivery delay is shorter as the number of replications of a message increases. This is because the chance that the relay nodes

having the replication meet the destination node is increased. However, buffer consumption of the relay nodes is larger as the replications of a message increase. Because of these properties, there is a trade-off between message delivery delay and buffer consumption.

So far, various DTN routing schemes have been proposed in order to resolve this trade-off and to transmit information effectively. Existing schemes are classified into several communication models. An example of these models is one-to-one communication models based on the host address. This communication model requires the name resolution between a source and a destination. However, the name resolution based on the host address is difficult in a network environment that consists of only wireless terminals such as DTN. Other examples of the models are the information dissemination-based communication model and information collection-based communication model for targeting all users. These models can be realized without the name resolution between a source and a destination. However, communication between the specified nodes is not possible. Therefore, in the DTN routing based on the existing communication models, each user in the network cannot select and get the information they wants.

In this paper, we present a DTN routing scheme based on the publish/subscribe model [3] that enables flexible communication by using topics of the information. The proposed scheme can communicate without checking each host address between sources and destinations because the name resolution is achieved on the basis of the topics of information. In addition, we proposed an algorithm that sort messages by the subscription lists and the contact condition of nodes in order to deliver to destinations with a short delay.

We compare the performance of the proposed scheme with that of existing schemes through simulations and show the effectiveness of the proposed scheme.

2 RELATED TECHNOLOGY

In this chapter, we discuss the functions of DTN routing and the publish/subscribe model as techniques related to our research. In addition, we discuss existing research on applying publish/subscribe model to DTN routing.

2.1 DTN Routing

Functions of DTN routing based on Store-Carry-Forward are classified into selecting relay nodes, selecting messages, and managing the buffer. In this section, we discuss the details of the three functions.

(1) Selecting relay nodes

With this function, each node selects the relay nodes to forward messages preferentially from several nodes within the communication range. The following is typical DTN routing schemes.

- Epidemic Routing [4]

Each node forwards the replications of the message to all nodes that it contacts. In this scheme, many replications of a message are generated, so the message delivery delay is short but buffer consumption is large.

- Two-Hop Forwarding [5]

Source nodes forward the replications of the message to all nodes that they contact, but relay nodes forward the replications of the message to only destination nodes. In this scheme, few replications of a message are generated, so the message delivery delay is long but buffer consumption is small.

- Spray and Wait [6]

In this scheme, the limit on the number of replications that can be generated from a message is set. After this limit is reached, each node waits to make contact with the destination nodes. In this scheme, the limited on the number of replications is set, so it is possible to control the trade-off between message delivery delay and buffer consumption.

- PRoPHET [7]

The relay nodes that are most likely to meet the destination node is selected from records of past communications of each node, and they receive the replication.

(2) Selecting Messages

With this function, the order to send messages is decided. It is expected that the session between nodes breaks down before each node forwards all messages to communication partners in DTN routing. Therefore, when they meet other nodes, they decide which message preferentially is forwarded from their buffer in order to improve the communication performance. Examples of the algorithm are FIFO (First In First Out), which messages in the order from oldest received time, and LIFO (Last In First Out), which messages in the order from newest received time.

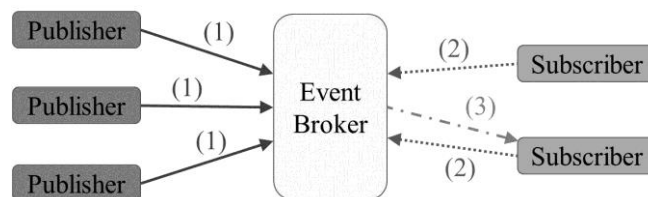
(3) Buffer Management

In this scheme, each node decides which message is removed in their buffer if the capacity of their buffer is exceeded because of an accumulation of messages. Examples of the algorithm are FIFO and LIFO as well as the function that select messages. In addition, another approach is to use recovery schemes that remove messages that are no longer needed after reaching destination nodes [8].

2.2 Publish/Subscribe Model

The publish/subscribe model [3] is a communication model that implements asynchrony between senders and receivers. The structure of this model is shown in Figure 1. It consists of three systems: publishers to send information, subscribers to receive information, and an event broker to relay information between publishers and subscribers. Publishers send the event broker the information they want to provide

in the network. Subscribers send the event broker requests for information they want to get from the network. An event broker checks the requests from subscribers and the information from publishers and sends subscribers the information that matches the requests.



Publisher: Sending the event broker information ... (1)

Subscriber: Sending the event broker requests ... (2)

Event Broker: Sending subscribers the information that matches the requests ... (3)

Figure 1: publish/subscribe model

Using a communication model based on the name resolution based on the host address is difficult in a network environment that consists of only wireless terminals such as DTN. In contrast, the publish/subscribe model can communicate without checking the host address between sources and destinations because the event broker achieves the name resolution based on the topics of information. Therefore, this model has the following advantages.

- It is not necessary that senders and receivers are synchronized temporally and geographically. Therefore, when the senders send messages, receivers do not need to be in the network.
- It is not necessary that senders and receivers notify their presence to each other because information is distributed on the basis of the content and topic of information.
- One-to-many communication, many-to-many is possible.

2.3 Publish/Subscribe-based DTN Routing

DTN assumes unstable network environments where frequent disconnections occur easily. The publish/subscribe model achieves asynchrony between the sender and the receiver. Therefore, these technologies are considered compatible, so the combination of them is attractive. In this section, we discuss related researches on DTN routing schemes based on the publish/subscribe model.

Kure proposed a routing scheme in which all nodes have the functions of the publishers, subscribers, and event brokers in DTNs constructed in disaster areas [9]. In the affected areas, all users who own a wireless terminal can be subscribers and publishers. In addition, special nodes that mediate between publishers and subscribers do not exist in the network all the time. Assuming these cases, each node processes communication on the basis of functions of subscribers, publishers, and event brokers. With the function of publishers, the nodes that generated messages forward the replications to all nodes in communication range at the present moment. With the function of subscribers, the nodes forward requests that contain the topics of information they

want to all nodes that they meet. In addition, the requests that they receive from other nodes in the past are forwarded. With the function of event brokers, if the nodes receiving the requests have messages corresponding to the request in their buffer, they propagate the messages by Epidemic Routing and deliver them to the node requesting. This routing scheme makes it possible for all the nodes in disaster areas to share the information with request and response. However, messages to be forwarded to the relay nodes are sorted by FIFO. Therefore, it is not always possible to deliver messages efficiently to all the destination nodes in the network.

Janico proposed a communication process (DPSP) for when two nodes meet [2]. The sequence of the communication based on DPSP is shown in Figure 2. When two nodes establish a session, they first exchange their subscription lists that contain the topics of information they want (1). Then each node builds a queue of the replications of the message from the local storage in order to forward the messages to the partner (2). After building the queues, the messages whose probability to be delivered is not improved when they are replicated to the partner are removed (3). This process has the effect of reducing the buffer usage of the relay nodes. Then, each node sorts the messages in their queue by their priority (4). After that, the nodes send the messages from the queues until the queues are empty or the session breaks down (5).

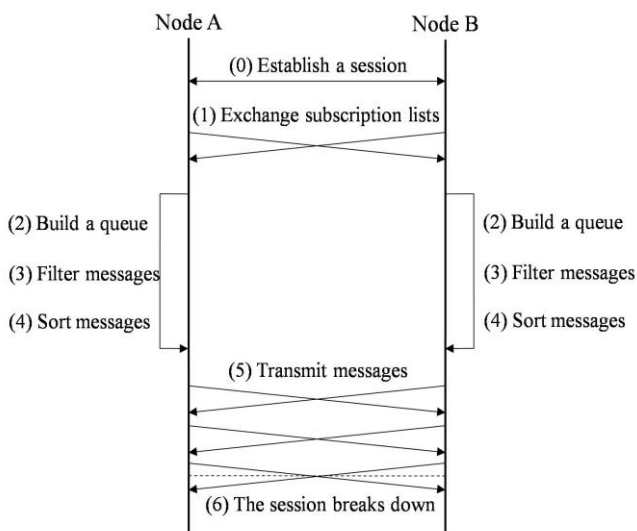


Figure 2: Communication process based on DPSP

In addition, Janico discussed several approaches for (3) filtering and (4) sorting the messages. Janico believed that message delivery rate and delivery delay are optimized by using these approaches properly depending on the network environment such as the message creation interval and the number of nodes. However, it is difficult for either approach to maintain consistently high communication performance without depending on the changes of the network environment. In addition, when the messages are sorted on the basis of the subscription lists, the communication performance of each topic may be uneven because of difference in the number of subscribers.

3 RESEARCH TASK

In this paper, we aim to establish a flexible system in which each user can select and get the information they want in a network environment that consists of only wireless terminals. Therefore, we need to discuss the efficiency of DTN routing based on the publish/subscribe model that communicates with topics of information. The challenge of this routing scheme is ensuring that the messages of each topic are delivered to subscribers with high probability and a short delay.

The capacity of each node's buffer and the time that it can communicate with other nodes are limited in a DTN environment. We need to improve the efficiency of message delivery under these constraints. In addition, our task is to reduce the bias in the communication performance of each topic without depending on the number of subscribers.

4 PROPOSED SCHEME

In this chapter, we propose a DTN routing scheme based on the publish/subscribe model detailed in section 2.3 and chapter 3. In the following, we discuss the assumed network environment and give a summary of the proposed scheme.

4.1 Assumed Environment

The proposed scheme is assumed to distribute information in a DTN that consists of only mobile nodes. Examples of the applications are services to deliver information on events, advertisements in the surrounding areas and news with high locality.

The nodes in the network are mobile phones, tablet devices, laptops, etc. All the nodes in the network can distribute information (messages) and receive it. Each message that is distributed in the network is belongs to a topic. Topic types are determined in advance, and new topics are not added. Each message is completed as one packet, and it is removed when TTL (Time To Live) expires.

4.2 Summary of Proposed Scheme

In the proposed scheme, all nodes have functions of publishers, subscribers, and event brokers that are based on the publish/subscribe model. Each node relays messages on the basis of the subscription lists and the contact condition of communication partners in order to deliver each message to the nodes subscribing to them. A use case diagram of the proposed scheme is shown in Figure 3.

Each node generates messages on topics that are specified as publishers. The header of a message is shown in Table 1. The nodes register the topics in which they have an interest in their subscription lists as subscribers. The elements of the subscription list are shown in Table 2. The nodes can set the order to receive messages and the priorities of the topics when registering the topics.

The nodes that receive messages and subscription lists relay the messages to the nodes that they meet as event brokers. The messages are sorted on the basis of the subscription list and contact condition of each node. In addition, each node delivers the messages matching the

topics in the communication partner's subscription lists as event brokers.

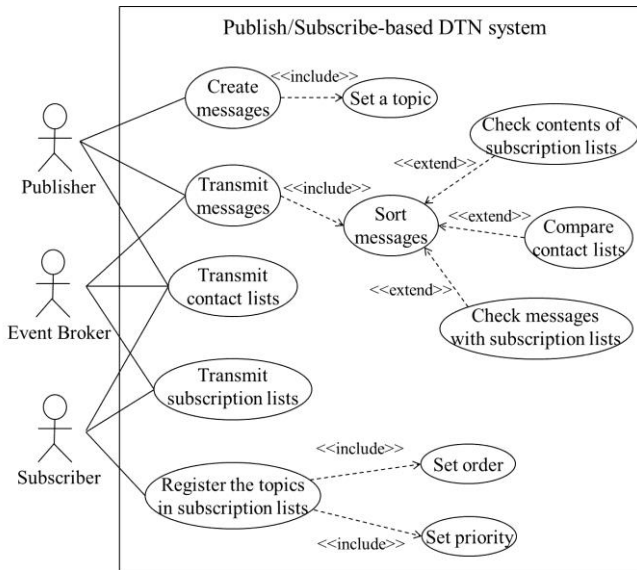


Figure 3: Use case diagram

Table 1: Header of a message

Element	Detail
<i>From</i>	Address of the node that generates a message
<i>TopicID</i>	Topic ID to which a message belongs
<i>MsgID</i>	Message ID, which contains a sequence number of a message to be counted independently for each node
<i>CreateTime</i>	Time that a message was generated
<i>ReceivedTime</i>	Time that a node received a message
<i>TTL</i>	TTL (Time To Live) of the message

Table 2: Elements of the subscription list

Element	Detail
<i>SubID</i>	Unique ID that a subscription list has
<i>TopicID</i>	Topic ID to which a message belongs
<i>Order</i>	Order in which to receive the message, which is selected from the ascending <i>CreateTime</i> of a message or descending
<i>Priority</i>	Priority of a topic
<i>ContactCnt</i>	Number of times that a node meets other nodes registered to the same topic

End to end communication is realized by each node processes communication on the basis of the above functions when it meets other nodes. The communication process between two nodes is shown in Figure 4.

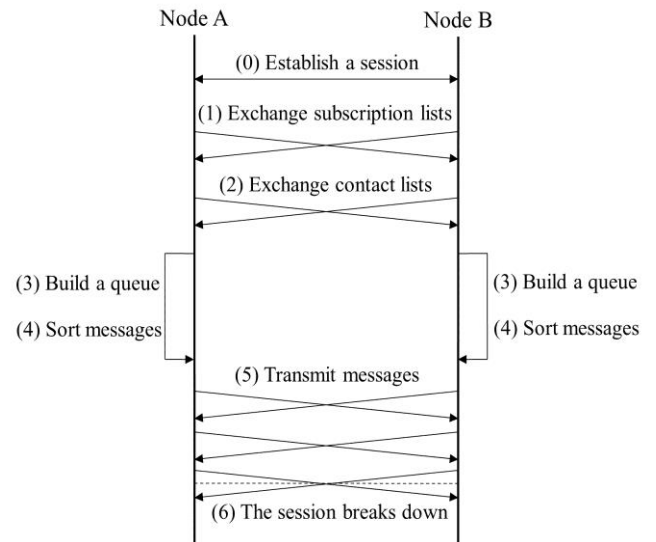


Figure 4: Communication process between two nodes

When two nodes meet, they first exchange their subscription lists (1). The subscription lists contain topics of information for each node and the nodes that it met in the past want.

After exchanging the subscription lists, they exchange their contact lists (2). The contact lists contain the average value of the contact time (i.e., how long each node is in contact with other nodes) and the average value of inter-contact time (i.e., the time between the end of a previous contact and the beginning of a new contact). They judge which of their contact condition is better by comparing their contact lists each other.

Then each node builds a queue of messages from the local storage to forward to the partner (3). After building the queue, the nodes sort the messages by the subscription lists and the contact lists (4). The details of this sorting algorithm are discussed in Section 4.3. After that, the nodes send the messages from the queues until the queues are empty or the session breaks down (5).

4.3 Sorting Messages

In a DTN environment, the session between nodes may break down before each node forwards all messages to a communication partner. Therefore, in the proposed scheme, the messages are sorted and transferred to the communication partner in the order of the highest priority.

The sorting of messages has two steps. In the first step, the priority of the topics to which the message belongs is determined (sorting topics). In the second step, the priority of messages that belong to the same topic is determined (sorting messages). Each node forwards the messages in the order that is determined through these steps.

In the proposed scheme, the messages are forwarded to the communication partner in three phases on the basis of the subscription lists and the contact lists. The procedure of forwarding the messages is shown in Figure 5.

(a) Forwarding the subscribed messages.

Each node forwards messages belonging to the topics that the communication partner registers. The sorting of topics is done in ascending the *Priority* that the partner sets when

registering the topic. The sorting of messages follows the *Order* that the partner sets when registering the topic.

(b) Relaying messages on the basis of the contact lists

Each node forwards the messages belonging to the topics in the subscription lists with the nodes that the partner met in the past. First, two nodes compare each contact condition by using the contact lists. Then, the messages with low reachability are relayed to the node whose contact condition is better, while the messages with high reachability are relayed to the node whose contact condition is worse. The messages of all topics are evenly propagated in the network by this process.

$V_{contact}$ is defined below as the indicator for evaluating the contact condition.

$$V_{contact} = \frac{T_{contact}}{T_{inter}} \quad (1)$$

In general, the number of messages that can be transferred during contact tends to increase as contact time $T_{contact}$ increases. The number of opportunities that for the node to contact with other nodes tends to increase as inter-contact time T_{inter} decreases. Therefore, a node whose $V_{contact}$ is large can communicate with many nodes for a long period of time. In this phase, as a result of comparing each node's $V_{contact}$, a node whose $V_{contact}$ is large relays messages with step (b-i), and a node whose $V_{contact}$ is small relays the messages with step (b-ii).

(b-i) A node whose $V_{contact}$ is large forwards messages with high delivery probability to a node whose $V_{contact}$ is small. The sorting of topics is done in descending *ContactCnt* in the partner's subscription lists. If some topics have the same *ContactCnt*, the topics are sorted in descending *Priority*. The sorting of messages follows FIFO.

(b-ii) A node whose $V_{contact}$ is small forwards messages with low delivery probability to a node whose $V_{contact}$ is large. The sorting of topics is done in ascending *ContactCnt* in the partner's subscription lists. If some topics have the same *ContactCnt*, the topics are sorted in ascending *Priority*. The sorting of messages follows FIFO.

(c) Forwarding the unregistered messages

The node forwards the messages that the partner's subscription lists are unregistered to. The messages are sorted by FIFO without sorting topics.

Both the messages of the popular topic and those of the unpopular topic are evenly propagated in the network by sorting the messages on the basis of the contact lists, not only the subscription lists. Therefore, the proposed scheme makes it possible for messages to be delivered to all subscribers more evenly than by sorting the messages on the basis of only the subscription lists.

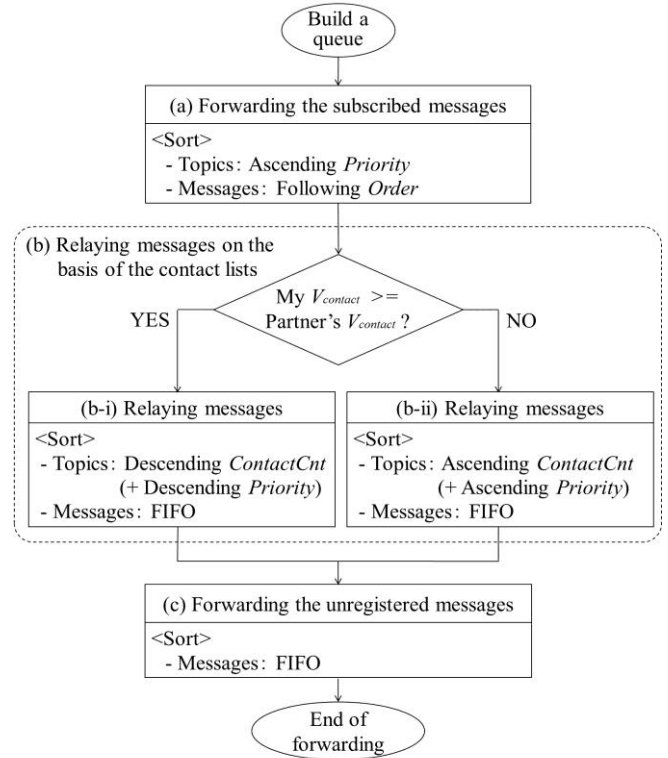


Figure 5: Procedure of forwarding messages

In addition, the messages are removed by FIFO if the capacity of their buffer is exceeded because of an accumulation of messages. This is because it is considered that the oldest message that each node has is relayed to other nodes sufficiently.

5 EXPERIMENTAL EVALUATION

5.1 Experiment Environment

We implemented the proposed scheme on the network simulator The ONE (The Opportunistic Network Environment Simulator) [11] and compared the performance of the proposed scheme with that of existing schemes in order to evaluate the effectiveness of the proposed scheme. The ONE is a simulator that was developed for evaluating of routing and application protocols in DTN environments.

The structure of the network that was used in this simulation is shown in Figure 6, and the simulation parameters that are shown in Table 3. There are three types of message topics that are distributed in the network: topics A, B, and C. There are 240 nodes that subscribe to topic A, and half of them move in Cluster P, and the other half move in Cluster S. There are 120 nodes that subscribe to topic B, and half of them move in Cluster Q, and the other half move in Cluster S. There are 40 nodes that subscribe to topic C, and half of them move in Cluster R, and the other half move in Cluster S.

Therefore, the nodes that subscribe to the same topic frequently contact with each other in Clusters P, Q, and R, and the nodes that subscribe to a different topic frequently contact with each other in Cluster S. In the simulation, it was evaluated whether messages belonging to each topic are

delivered to the nodes whose contact condition is different in the network.

A node is assumed pedestrian and moves by Random Waypoint. A message is assumed advertisements and news and is generated by a node selected randomly from all nodes. The message size is determined by a normal distribution when the message is generated. Messages are generated until 30 minutes before the end of the simulation time. A communication standard of a node is assumed Bluetooth.

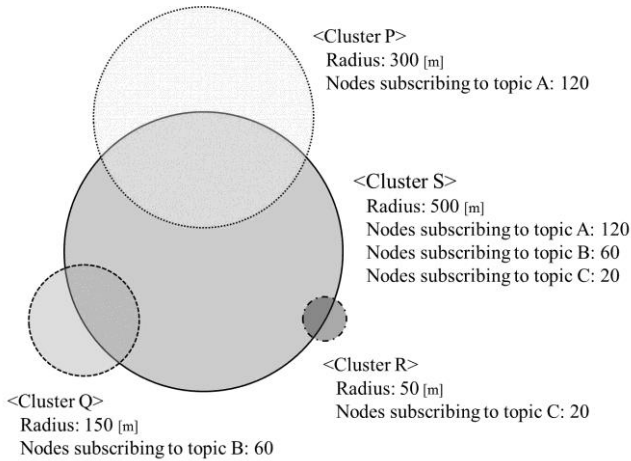


Figure 6: Structure of network

Table 3: Simulation parameters

Simulation time	12 [hour]
Moving speed of node	1.8 - 5.4 [km/h]
Wait time	0 - 120 [sec]
Communication range	10 [m]
Communication speed	250 [Kbps]
Buffer size	50 [MB]
Message size	0.5 - 3.0 [MB]
Message creation time	11.5 [hour]
TTL	120 [min]

5.2 Evaluation Policy

In the experiment, the proposed scheme is compared with the existing schemes in terms of average message delivery rate and average delivery delays as the message creation interval changed. In addition, the standard deviation of each scheme was compared in order to evaluate the effect that the number of subscribers has on the communication performance. The definition of each evaluation indicator is the following.

- Message delivery rate (Avg.)
The rate of the subscribing nodes who received each generated message
- Message delivery delay (Avg.)
The average time between the creation and arrival of the message that was delivered to the subscribing node

The existing schemes compared with the proposed scheme are Epidemic Routing, Two-Hop Forwarding, and Subscription-based Routing (SBR). The SBR is basically the same mechanism as the proposed scheme, but the messages are sorted on the basis of only the subscription lists without

the contact lists. Therefore, the messages are forwarded to all nodes in the order of (a), (b-i), and (c) in Figure 5. In addition, the messages are sorted by FIFO in Epidemic Routing and Two-Hop Forwarding.

5.3 Results of Experiment and Discussion

The average message delivery rate is shown in Figure 7, and the average message delivery delay is shown in Figure 8.

As shown in Figure 7, the average message delivery rate of the proposed scheme was always more than 99% in parallel with Epidemic Routing and SBR. It is considered that the DTN routing scheme based on the publish/subscribe model can reliably forward messages to each subscribing node through this result.

As shown in Figure 8, the average message delivery delay of the proposed scheme was about 80 - 100 seconds shorter than that of Epidemic Routing and SBR, and about 400 seconds shorter than that of Two-Hop Forwarding. This is because messages with a lower priority were propagated actively by the nodes with good contact condition.

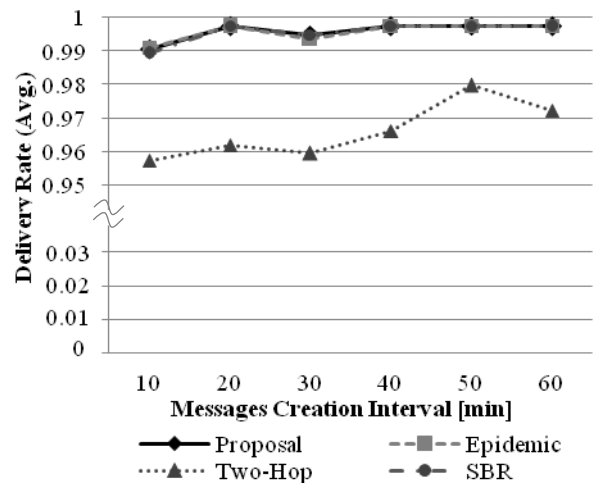


Figure 7: Message delivery rate (Avg.)

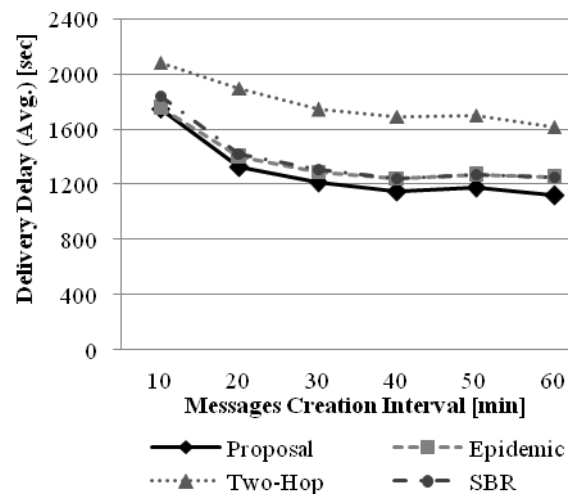


Figure 8: Message delivery delay (Avg.)

In addition, the standard deviation of the message delivery delay of the proposed scheme, Epidemic Routing, and SBR are shown in Figure 9. The standard deviation of the proposed scheme was smaller than that of other schemes as the message creation interval decreased. It is considered that the proposed scheme prevents an increase in the message delivery delay due to the number of subscribing nodes because each message is relayed in accordance with the priorities and the contact condition even if the message is generated frequently.

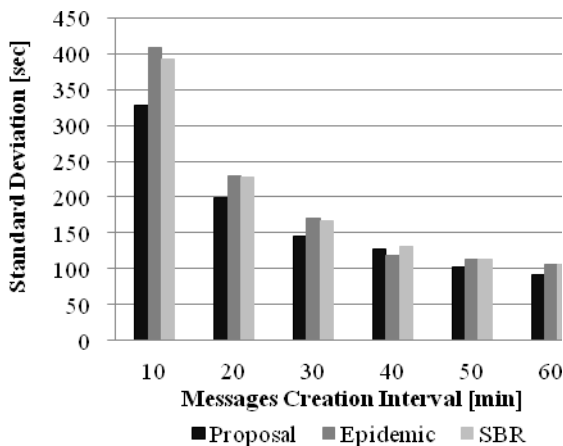


Figure 9: Standard deviation of message delivery delay

6 CONCLUSION

In this paper, we proposed a DTN routing scheme based on the publish/subscribe model with the aim of establishing a flexible system with which each user can select and get the information they want in the network environment that consists of only wireless terminals. We compared the performance of the proposed scheme with that of existing schemes on the network simulator The ONE in order to evaluate the effectiveness of the proposed scheme. Through the results of the experiment, the proposed scheme was confirmed to deliver messages to subscribing nodes with high probability. In addition, message delivery delay of the proposed scheme was about 80 - 100 seconds shorter than that of the existing schemes, and its dependence on the number of subscribing nodes was low.

It is considered that the effectiveness of the proposed scheme increased as the simulation map had a high characteristic of the mobility of nodes, because the messages are sorted by the contact condition of the nodes, although the mobility model of the nodes was set for Random Waypoint in this experiment. Therefore, we will implement a map and a mobility model that are closer to the real world and evaluate the effectiveness of the proposed scheme. In addition, our challenge is also to study approaches that reduce buffer consumption.

REFERENCES

[1] S. Farrell and V. Cahill, "Delay and Disruption Tolerant Networking," Artech House, 2006.

- [2] V. Cerf, S. Burleigh, A. Hooke, L. Torgerson, R. Durst, K. Scott, K. Fall, and H. Weiss, "Delay tolerant network architecture," IETF RFC 4838, 2007.
- [3] P. Th. Eugster, P. Felber, R. Guerraoui, and A-M. Kermarrec, "The Many Faces of Publish/Subscribe," ACM Computing Surveys, Vol.35 No.2 pp.114-131, 2003.
- [4] A. Vahdat and D. Becker, "Epidemic routing for partially-connected ad hoc networks," Duke Technical Report, CS-2000-06, 2000.
- [5] Z. Zhang, "Routing in intermittently connected mobile ad hoc networks and delay tolerant networks: Overview and challenges," IEEE Communications Surveys, Vol.8, pp.24-37, 2006.
- [6] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Spray and Wait: An efficient routing scheme for intermittently connected mobile networks," Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking, pp.252-259, 2005.
- [7] Z. Haas and T. Small, "A new networking model for biological applications of ad hoc sensor networks," IEEE/ACM Transactions on Networking, Vol.14, No.1, pp.27-40, 2006.
- [8] K. Kure and H. Esaki, "Pub/Sub based Information Acquisition System over Disaster Area DTN," Proceedings of the DICOMO2010 Symposium, pp.1352-1359, 2010. (*in Japanese*)
- [9] J. Greifenberg and D. Kutscher, "Efficient Publish/Subscribe-Based Multicast for Opportunistic Networking with Self-Organized Resource Utilization," Proceedings of IEEE International Workshop on Opportunistic Networking (WON-2008), pp.1708-1704, 2008.
- [10] A. Lindgren, A. Doria, and O. Schelén, "Probabilistic routing in intermittently connected networks," Lecture Notes in Computer Science, pp.239-254, 2004.
- [11] A. Keränen, J. Ott, and T. Kärkkäinen, "The ONE Simulator for DTN Protocol Evaluation," Proceedings of the 2nd International Conference on Simulation Tools and Techniques (SIMUTools'09), p.10, 2009.

Sales and marketing support for BtoB using Web form

Hiroshi Horikawa

Planning Department, Mitsubishi Electric Information Technology Corporation, Japan
horikawa-hi@mdit.co.jp

Abstract -This paper explains Web form of an inquiry, Web form of reports download, and seminars registration form as an aid of the cognitive improvement in the importance of the website for BtoB as an action tool (marketing tool) in a website. I describe the change in the number of accesses, the sales and marketing support effect, and increase in efficiency and real-time nature as an examination of Web forms. In BtoB, since the user is performing the input as part of work from these things on behalf of the organization to which the user belongs, it is rare to be interrupted on the way, and can be observed that Web form does not become a factor which reduces the number of accesses. It is easy to manage the data inputted into form by RDB (Relational DataBase), since the merit of the form can distinguish a mail address and an address, the diversion to an e-zine or direct mail becomes easy, and it also tends to take statistical information. For this reason, in every form, the efficiency of the work of data reduction can be increased.

Keywords: Web, Web Form, BtoB, Sales and Marketing, Marketing Tool, Action Tool.

1 INTRODUCTION

This paper explains sales and marketing support of the website of Mitsubishi Electric Information Technology Corporation (MDIT). MDIT is a company that performs manufacture, sales, and maintenance for computers of BtoB. BtoB (Business to Business) refers to the business for corporations, and, on the other hand, BtoC (Business to Customer) refers to the business for individual customers. When only the number of accesses is compared simply, usually BtoB is far less than BtoC [2]. Generally, in Japan, the door-to-door sales with personal business in BtoB are in the mainstream. Also making the selling structure using the Internet is behind the times. There are not many BtoB companies which recognize and understand the importance of website. This paper explains Web form of an inquiry, Web form of reports download, and seminars registration form as an aid of the cognitive improvement in the importance of the website for BtoB as an action tool (marketing tool) in the website. An action tool here refers to the dynamic page on the website made by programming languages, such as Java.

2 SALES AND MARKETING SUPPORT FUNCTION OF WEBSITE FOR BTOB

In 2009, I was ordered carrying out sales expansion using a website from the company. Then, I investigated in-

company needs while investigating the trend of the improvement direction of a website. The characteristic of a website is outlined to 2.1 and the improvement means of a site is outlined to 2.2. In response to the request of Web form from the sales and marketing division in the company, I developed Web forms in 2010. Since the first proposition was "carrying out sales expansion using a website", activity here is evaluated as (2) BtoB site of 2.1.

However, the thing concerning (1) Company brand of 2.1 was also contained in the request of a sales and marketing divisions. In improvement of a site, carrying out all the improvement indicated to 2.2 is being continued repeatedly. The action tool in the website for BtoB explained in this paper belongs to (2).

2.1 Characteristic of Websites

Kiga[3] divided the website into three from the characteristic, as shown below.

(1) Company brand sites

Generally a corporate section manages this kind of site. The goal will become improvement in company brand awareness. This goal is long-term and is not necessarily easy to measure. Elsewhere, an education sites also have these characteristics.

(2) BtoB sites

Generally an operating department manages this kind of site. The goal can be made into the item which can be measured in the short term. For example, the number of inquiries, the number of orders received. Elsewhere, product brand sites and campaign sites also have these characteristics.

(3) EC (Electronic Commerce) sites

EC site can be judged a value more directly, and can consider a success as trade. Elsewhere, WebEDI(Electronic Data Interchange) sites also have these characteristics.

2.2 Improvement Means of Sites

Tomaru[4] divided the improvement means of a site into three as shown below. In addition, these improvements are carried out repeatedly.

(1) Access analysis

First, access analysis is conducted and the directivity of a Web improvement is defined.

(2) Contents enriching

Next, contents are enriched because increase the amount of information for the purpose of stay and conversion of access.

(3) SEO (Search Engine Optimization) solutions, Web advertisements

Furthermore, visitors are gathered using the measures against search engines (setup of keywords), or Web advertisements.

3 THE SITE MAP OF MDIT

The website of MDIT has taken 3 general hierarchy organizations of the Web server, the application server, and DB server, as shown in Figure 1 server block diagram.

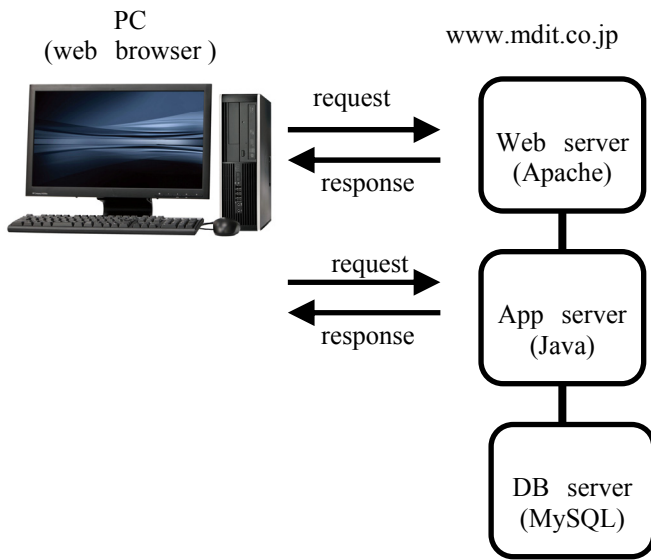


Figure 1: Server block diagram.

The site map of MDIT (extract) is shown in Figure 2. The top page of MDIT is shown in Figure 3. From the top page, visitors can move to the page of Hot news, News release, Corporate information, Job opportunities, Inquiry and introduction of products and services which MDIT offers.

Top page of MDIT

Products and services list (About 50 kinds)

-
-
-
- Data analysis platform “AnalyticMart”
 - Products composition
 - Brochures / Customer reports
 - Reports download (form)
 - Services
 - Inquiry
 - Personal information policy
 - Inquiry (form)
 - Seminars
 - Seminar list
 - Seminar Information
 - Personal information policy
 - seminar registration (form)

Hot news, News release, Corporate information, Job opportunities and Inquiry

Figure 2: The site map of MDIT (extract).



Figure 3: The top page of MDIT (only in Japanese).

The page of Data analysis platform “AnalyticMart” which is one of about 50 kinds of products and the services is shown in Figure 4. Visitors can move to the page of Products composition, Brochures / Customer reports, Services, Inquiry, and Seminars further from the page of Data analysis platform “AnalyticMart”.



Figure 4: Product introduction page (example: Data analysis platform “AnalyticMart”).

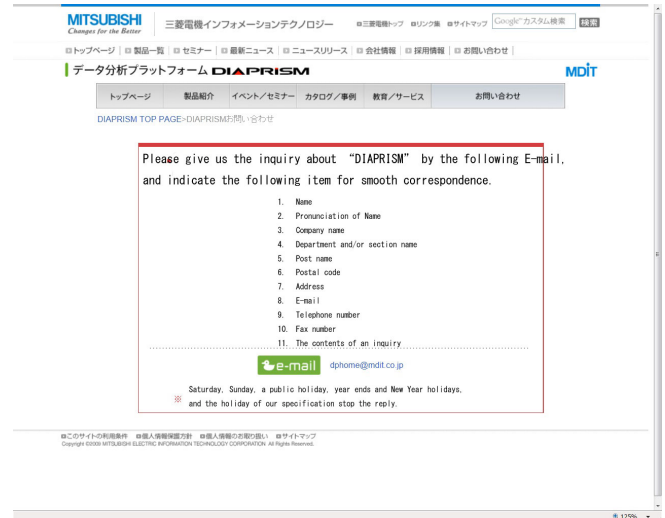


Figure 5: The inquiry page (Plain text).

4 WEB FORM

Web form of inquiry (4.1), Web form of reports download (4.2), and seminars registration form (4.3) described in this paper explain the page displayed, respectively when Inquiry, Brochures / Customer reports, or Seminars is chosen. Users push Inquiry button in order to send an inquiry to MDIT, users push Brochures / Customer reports button in order to download PDF of Brochures or customer reports, and users push Seminars button in order to write a registrant to the seminar about products and services.

4.1 Web Form of Inquiry

I developed Web form of inquiry in response to the request from in the company. Figure 5 shows the page which presents the information about the inquiry before setting Web form. Before setting Web form, Web of MDIT did not have an action tool, and then the website of MDIT has taken only 1 hierarchy organization of the Web server. Since the Web server did not have an action tool, MDIT had eased the burden of security countermeasures when managing the server. And the security merit was the reason why my section did not give a server an action tool. However, since members of sales and marketing divisions desired Web form of the inquiry, my section installed the action tool in the server. The website of MDIT became three hierarchy organizations then. Figure 6 shows Web form of inquiry.

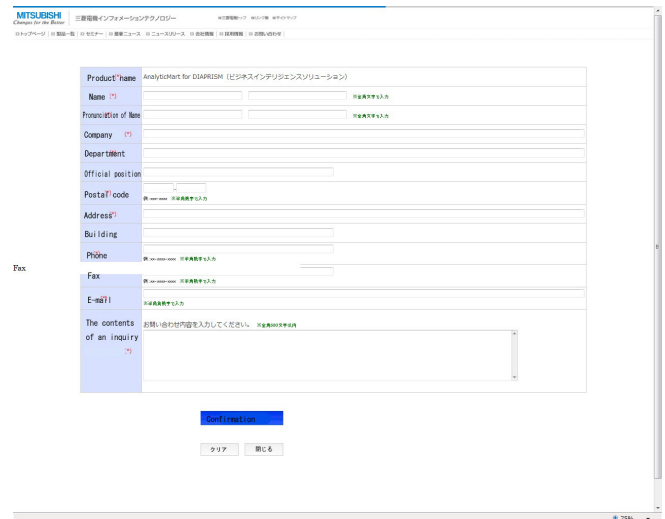
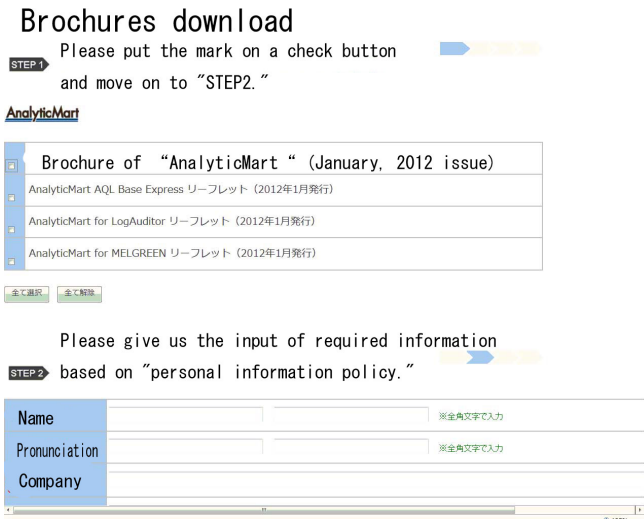


Figure 6: Web form of inquiry.

4.2 Form of reports download (Potential customer information acquisition)

Before setting form of reports download, we provide the user with PDF of Brochures and customer reports freely. However, we set the form into which visitors input a visitor's personal information before reports download, and we provide these input information as a potential customer's information to the sales and marketing divisions. Figure 7 shows the report download form.



After choosing the reports to download, visitors input a variety of information. After the input frame "Company name" of a screen, the input frame an "Department and/or section name", a "postal code", the "address", "PHONE", "E-mail", and a "Remarks" continues.

Figure 7: The reports download form.

I prepared the program for form generation, in order to generate reports download forms in a site. An operator only specifies "the logo of product", a "report name", "an actual report", and "the mail address for in-house use which a program mails at the time of download" to the program, and can generate report download form.

4.3 Seminar Registration Form

The marketing activities using a seminar are effective as a method of gaining a potential customer. Figure 8 shows seminar registration form. A seminar registration form consists of seminar information, a questionnaire, and a visitor information input.

The questionnaire item in middle of the screen "Who guided this seminar" can be changed for each seminar. Figure 9 shows the screen that sets up seminar registration form. Operators set up seminar information and a questionnaire. And, operators have to specify the mail addresses that tell the arrival of registration, the e-mail signature of a registration confirmation mail and the URL of seminar information other than a setup of seminar information and a questionnaire. When there was no seminar registration form, we had obtained seminar registration by e-mail or FAX.



Figure 8: Seminar registration form.



Figure 9: The screen that sets up seminar registration form.

5 AN EXAMINATION OF WEB FORM

I describe the sales and marketing support effect, the change in the number of accesses, and increase in efficiency and real-time nature as an examination of Web form.

5.1 Sales and Marketing Support Effect

This paper describes three sorts of Web form. Web form of report download has the sales and marketing support effect clearly (Table 1, Table 2). Before, even if the information of the user who downloaded the report analyzed the log of Web, it understood only the user's IP address (in addition, a company name may be able to become clear from an IP address). At the time of download, potential customer information can be economically obtained now because user information comes to hand. On the other hand, there is no

change in the information on sales and marketing which can obtain an inquiry and seminar registration.

Table.1 Number of Personal Information Acquisition

FY	Visits	Acquirement	Conversion rate
2009	252,000	653	0.3%
2010	Web Forms are developed		
2011	263,000	1,592	0.6%
2012	226,000	1,323	0.6%

Table.2 Number according to Form

FY	Acquirement	Inquiry	Download	Seminar
2009	653	333	0	320
2010	Web Forms are developed			
2011	1,592	315	963	314
2012	1,323	291	712	320

registration form automates no vacancy management, and increases the efficiency of seminar receptionist work.

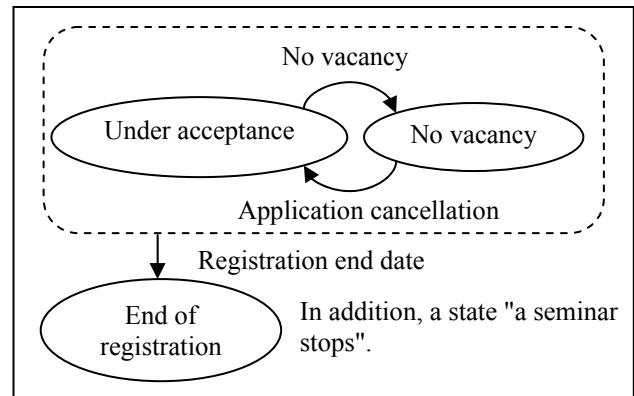


Figure 10: The change state of seminar registration.

5.2 Number of Accesses

There was no big change in the number of accesses before setting form, and after setting form (Table 3). As Web form of inquiry was requested by sales and marketing divisions in the company, we expected that the number of accesses would increase by Web form. However, number of accesses did not increase. On the contrary, as a possibility of interrupting an input triggered by the inconvenience which must be indicated in a fixed form became high, we expected that the number of accesses became fewer. However, there was no big change in the number of accesses. Since users are performing the input as a part of work from these things on behalf of the organization to which the user belongs, there are rare to be interrupted on the way. And we can observe that Web form does not become a factor which reduces the number of accesses.

Table.3 Changes of Access (Conversion)

FY	Visits	Inquiry	Seminar
2009	252,000	333	320
2010	Web Forms are developed		
Average (2011-2012)	245,000	303	317
Change	-3%	-9%	-1%

5.3 Increase in Efficiency and Real-Time Nature

The seminar application form has an effect of reducing the work in MDIT. By introducing seminar registration form, when the visitor registered, the visitor can do a check by the E-mail, and MDIT could know arrival of registration by the E-mail. As the change state of seminar registration of Figure 10 shows, the program checks no vacancy in real time, and if it becomes no vacancy, it will end registration. However, if there is application cancellation also in the state of no vacancy, acceptance will restart in real time. Compared with the seminar application using FAX or E-mail, seminar

It is easy to manage the data inputted into form by RDB, since the merit of form can distinguish a mail address and an address, the diversion to an e-zine or direct mail becomes easy, and it also tends to take statistical information. For this reason, in every form, the efficiency of the work of data reduction can be increased.

6 CONCLUSION

6.1 From partial optimization to overall optimization

Three sorts of action tools described in this paper also have the sales and marketing support effect. However, there were few population parameters of the conclusion number and we were not able to show a basis to the extent that it becomes significant statistically. According to Hanzawa [2], the effect of the site at the time of the merchandise purchase of BtoB is 26.5%, and the effect of the site at the time of the merchandise purchase of BtoC is 7.6%. That means the BtoB of the role of Web is larger than BtoC. As the background, we assume that the visitor of BtoB judges synthetically various sources of information containing a website as work, on the other hand, the visitor of BtoC purchases in the shop, without investigating the contents of goods by Web. In addition, in BtoB, it has possibility that only one access will lead to a big-ticket order received, and worth of access is high. There is a gap in judgment of worth of Web by the customer and a supplier, and it sees low in the supplier side. For this reason, we are not necessary to evaluate the Web independent sales and marketing support effect, but we need to estimate the whole sales and marketing process including the sales and marketing support by Web. For example, a user company name is searched for from an IP address, and the customer who received the proposal letter and the estimate from MDIT is analyzed through Web access. We need to consider optimization in the whole sales and marketing process includes Web.

6.2 From Information Publishing Site for Sales Partners to WebEDI

The sales partner-oriented information publishing site was established also by MDIT as BtoC has a member site. The purpose to install the information site for sales partners is strengthening of the existing sales channel. The main roles of the information site for sales partners are offer of the newest price information, and offer of products and services information concerning sale. The action tools in this site are a function which carries out visible control of the model which can be treated from the corporate information to which a user belongs, and the function to adjust an offer price from the past volume of sales. The model information, which controls a display for each user, is based on the sales partner contract. Control of a display is managed as an attribute of user information (refer to Figure 11). From the past volume of sales, as a function to adjust an offer price, several price lists are prepared, and it adjusts manually (refer to Figure 12). Before establishment of this partner's site, E-mail was used for transfer of price information or products and services information. After establishment of this partner's site, the work and cost concerning information publishing have been reduced for MDIT (Table 4). And the partner can peruse information in real time. We would like to try WebEDI and to go as strengthening of the further existing sales channels by Web in order that we may incorporate MDIT products and services information into the system of partners or customers.

Operators choose the rank according to model with a pull down menu for each company of users. When a user's download is required, the price list of the rank of the number specified here will be downloaded. If operators specify a model rank as "0 (standard)", the display item of "the price list for partner" will disappear from a partner's (user) screen. An operator's specification of "please choose" will vanish all the corresponding model buttons from a user's screen (it becomes a model which cannot treat the specified model).

Figure 11: User information management.



Operators upload the standard price table which are called the rank 0 and three price lists for partners (rank A-C), after setting up the date of issue, a title, a classification, and the file name at the time of download. The price lists for a partner become partner's cost price. A cost price will also become cheap if a rank increases. Operators specify the rank of a certain user's price by the user management previously shown in Figure 11.

Figure 12: Price information management.

Table.4 Cost reduction
(Information Publishing Site for Sales Partners)

$$5\text{minutes} \times 600 \text{ Letters/year} \times 10\text{Branch} = 3\text{man-month}$$

7 CLOSING REMARKS

I am wishing the following two things. Standardization should move on and anyone can use form easily as a mechanism [12] [13] [14]. Web Service should spread, there are no collateral conditions, such as hardware, and anyone can use form easily.

APPENDIX OUTLINE OF MDIT

Since the state of a website is affected from the state of a company, the outline about MDIT is shown below.

Yearly turnover: 31,900 million yen (320 million USD, the 2011 fiscal year).

The number of employees: 953 persons (the 2011 fiscal year)

The number of products models: 46 kinds

Custom-made and general-purpose ratio of products: 76% of the Products are general-purpose products.

APPENDIX WEB ANALYTICS (FY 2012)

Page View	: 878,000
Visits(Session)	: 226,000
Unique User	: 148,000
Average Session Duration	: 00:01:59

COPYRIGHT

The copyright of all the Web pages published in this paper is reserved by MDIT.

REFERENCES

- [1] Hiroshi Horikawa, The goods order system by a Personal Digital Assistant, the Institute of Electronics, Information and Communication Engineers(IEICE), Vol 80, No 4, pp390-394 (1997).
- [2] Akihiro Hanzawa, The importance of the website in BtoB business, Trade advertising, and Vol 43, No.10, and pp10-16 (2011).
- [3] Takashi Kiga, "Web master inquiry" is serialized. -As actual as the spot of the Web practical use in a B to B company -, Trade advertising, Vol 43, No 6, and pp10-14 (2011).
- [4] Tomaru Ryuzo, Let's have Web marketing interlocked with a sales process, That's eigyo, Vol 1, pp110-113 (2006).
- [5] Advertisement meeting, Social-media practical use / company brand, an education site / BtoB site / campaign site / EC site / brand site, An advertisement meeting, the volume 805, pp.18-36 (2011).
- [6] Advertisement meeting, It leads to an order received! The newest process of digital manipulation -Strategic practical use of WEB!! Evolving BtoB marketing-, An advertisement meeting, Vol 855, pp18-47 (2013).
- [7] Makoto Yuasa, The future communication activities considered from a user's information gathering course, Advertisement meeting, Vol 825, pp84-87 (2011).
- [8] Kayo Murakami, Naomi Kimura. Hideki Asahi, Yasushi Iwami and Masashi Tozawa, B to B company of Japan-digital-Possibility of marketing practical use -, Advertisement meeting, Vol 806, pp98-100 (2011).
- [9] Advertisement meeting, The WEB site of BtoB - The opportunity of new acquisition is lost? -, Advertisement meeting, Vol 781, pp100-103 (2010).
- [10] Hiroki Sagawa, The support example of a diagnostic checker, Corporate diagnosis, Vol 57, No 10, pp30-33 (2010).
- [11] Nikkei computer, Seven keywords which read and solve BtoB - From Web-EDI to a business model patent -, Nikkei computer, Vol 498, pp201-215 (2000).
- [12] Dubinko, Micah and et.al, , Xforms 1.0, W3C Recommendation (2003).
- [13] W3C, Extensible Markup Language (XML) 1.0(Fifth Edition),(online), available from <http://www.w3.org/TR/REC-xml> (2008).
- [14] Hideki Kojima, Concrete approach to the electronic application by Web/XML / electronic form practical use, Japan Society of Information Knowledge, Vol 11, No 3, pp.21-26 (2001).

Session 2:

Education and Business

(Chair: Tomoo Inoue)

Proposal and Evaluation of Collaborative Attribute Method in Text Recommender Systems for E-Learners

Yuji Wada^{*}, Takuya Segawa^{*}, Jun Sawamoto^{**} and Hiroyuki Sato^{**}

^{*}Tokyo Denki University, Japan

yujiwada@mail.dendai.ac.jp

^{**}Faculty of Informatics, Shizuoka University, Japan

^{***}Iwate Prefectural University, Japan

sawamoto@iwate-pu.ac.jp

Abstract -E-learning is used in various places. However, many systems do not show advantages, such as online exams, and simply enumerate the teaching material, etc. In our An Individual Reviewing System (abbreviated AIRS), contents of each user are optimized according to recommendations using Collaborative Filtering (what we call CF). This system multiplies the load to the user by smoothly improving study efficiency. However, this CF method has disadvantages in that if insufficient data is available, recommendations may show poor accuracy. To overcome that disadvantage, we suggest Attribute Correlation method that uses metadata which users have. We experimented with this method, and the result was not good. A new approach (called collaborative attribute method) is proposed to address the problems identified and show the experimental results.

Keywords: Recommender System; Web Digital Texts; E-Learning; Cold-Start Problem.

1 INTRODUCTION

E-learning, in which students can learn anywhere, at any time, has been coming into broader use in universities, corporate training and other settings. Many existing systems, however, simply make teaching materials available and conduct online testing, without providing the full range of unique learning advantages available through e-Learning. With an individualized reviewing system (AIRS), provision of content is tailored to the specific learner, as described in Wada, Matsuzawa, Yamaguchi, and Dohi [1]. This system uses an algorithm that helps students learn efficiently, based on the student's own historical data and the historical data of other learners, as discussed in Wada, Hamadume, Dohi, and Sawamoto[2]. One disadvantage, however, is that it cannot handle recommendations before any historical data have been accumulated. To solve this, we proposed an attribute correlation method using the background data of the user, and evaluated the usefulness of this approach in Wada, Segawa, Sawamoto and Sato[3]. The results, however, did not show this method to be particularly useful. So, we have proposed a different approach.

- We proposed an attribute correlation method using the background data of the user.
- We tested subjects using the proposed method, and evaluated the results.
- Is the proposed method effective?
- We proposed a different method, after considering improvements to the proposed method.

2 RELATED RESEARCH

Research in systems that anticipate user preferences and recommend contents is currently advancing, with a number of Web services using this approach. For instance, with the EC services used by Amazon[4], products are recommended that are likely to appeal to the user, based on the user's product page viewing history, purchasing history and other data. Many of these systems use collaborative filtering (CF), as shown in Resnick, Iacovou, Suchak, Bergstorm, and Riedl [5]. In terms of education, however, research in the use of CF as opposed to education based on classroom lectures and other realistic environments is being conducted, as discussed in Kitamura[6], but there are few cases in which this has actually been incorporated into e-Learning systems. With AIRS, learning content is recommended to the learner. With CF, however, a problem called Cold-Start exists, in which the user has to use the contents to some extent, or no history can be obtained, and this makes it impossible to provide recommendations with a high level of accuracy, as described in Schein, Popescul, Unger, and Pennock[7]. This poses a drawback for users who want to use the system to solve questions in content learned through lectures and other means, or to review content already acquired. The research presented here proposes the attribute correlation method, which focuses on the Cold-Start problem.

3 COLLABORATIVE LEARNING RECOMMENDATIONS

Collaborative learning recommendations are recommendations carried out through the same procedure as CF. Hereafter the user will be referred to as the "learner", and the historical data as "learning history". The procedure for making collaborative learning recommendations comprises the following sequence of steps.

This research was supported by one Grant-in-Aid for Scientific Research C (Subject No. 21500908), and is currently supported by the other Grant-in-Aid for Scientific Research C (Subject No. 24500122).

3.1 Extraction of Similar Learners

Other learners who have preferences similar to those of the learner for whom contents are to be recommended are extracted as “similar learners”. A database of the learning histories of learners is compiled, and correlations are drawn between learners based on that database, with learners being sorted in sequence based on the size of the correlation coefficient. Higher-order learners with a particularly large correlation are extracted as similar learners.

3.2 Extraction of Recommendation Contents

The actual content to be recommended is extracted from among the learners extracted as similar learners. The learning histories of similar learners are used to identify difficulties encountered by those persons, and analogies are drawn based on the way that those difficulties were overcome in order to extract relevant content.

3.3 Presentation of Recommendation Results

The extracted content is presented to the user via the system. This involves the system interface, and will not be addressed here.

4 ATTRIBUTE CORRELATION METHOD

As described above, collaborative learning recommendations are formulated by selecting recommended content based on the history of the learner. For this reason, similar learners cannot easily be extracted for learners who do not already have a learning history, or learners for whom a certain level of learning history has not been compiled (hereafter, we will call these “new learners”). As a result, it will not be possible to present highly accurate recommendation results. Given this, we propose a method of extraction in which background data for new learners is compiled and treated as attribute data, and learners with attribute data similar to that of the learner for whom recommendations are being provided are extracted as persons with similar attributes.

4.1 Overview

A primary reason for the Cold-Start problem that occurs in the collaborative learning recommendation method is that new learners do not have extensive histories, making it difficult to identify similar learners, as described in Section 3.2. In other words, this problem could possibly be solved if correlations between new learners and existing learners could be evaluated by other means. Figure 1 shows an overall flowchart incorporating the proposed method.

4.2 Attribute Data

Attribute data are acquired from meta-data, for example, age, sex, hobbies and preferences, strong subjects, weak subjects, and other personal data. This data is certain to be available for new learners, even if they do not have a learning history.

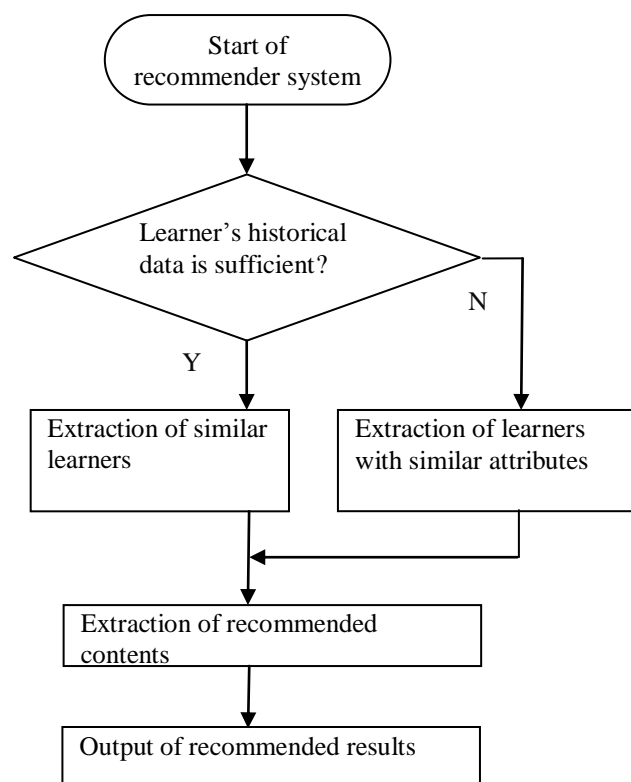


Figure 1: Attribute correlation method flow chart.

4.3 Systematization of Attribute Data

In attribute data, there is relevance among data items. For example, no relevance can be identified in a high school education between writing and physics, but a certain degree of relevance can be found between items that are both in a science curriculum, such as mathematics IA and physics. Systematizing attribute data within itself and expressing it is believed to be a necessary step, the reason being that one can envision that there will be little attribute data that can be compared to the learning history and used.

With learning attributes, taking, for instance, a high school education as an example, coursework subjects are classified into root nodes, with science-based classes and literature-based classes as sub-nodes. These sub-nodes are further classified into generalized coursework classifications. Even more detailed names and definitions of classes are provided at the next layer, and a hierarchical structure is created. Moving further down the hierarchical layers, data become more specific, and thus carry greater weight as information. This weight can be expressed in terms of points: the first layer directly beneath the root node is counted as 0.5 points; and underlying layers are counted as 1, 2, and 4 points respectively, so that each layer has double the weight of the layer just above. This is done to increase the estimated value of the deeper layers. Figure 2 shows an example of the systematization of attribute data pertaining to learning. Here, only those types of attributes necessary for the evaluation, such as “learning” and “occupation”, are created.

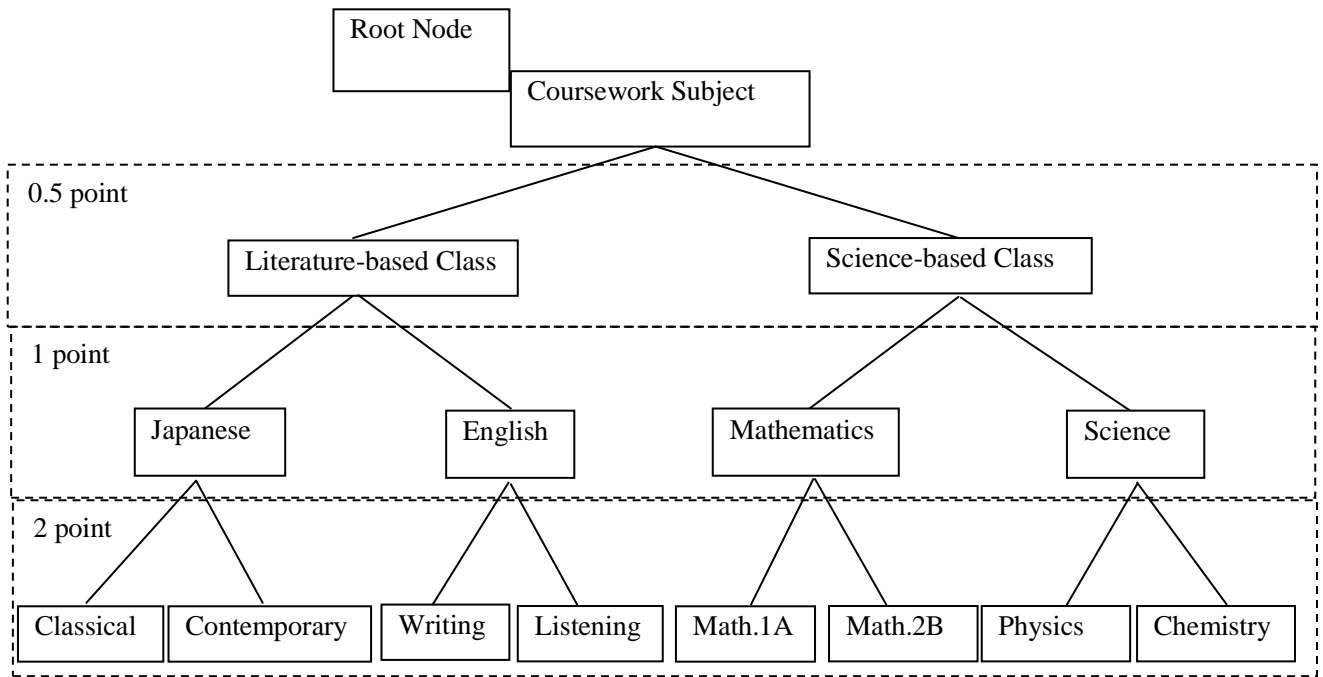


Figure 2: Hierarchy of attribute data.

4.4 Extraction of the Degree of Attribute Data Similarity and Users with Similar Attributes

Conformances of attributes between new learners and all other learners are compared, and scores of all of the attributes are added together. A ranking is then created, with the highest scores at the top, and learners with particularly high conformance values are taken as learners with similar attributes. In the example shown in Table 1, Learner N is strong in the subject of physics, and thus has information in science and in science-category classes, which are upper-level nodes. Learner X matches completely, so 2 points are assigned, while Learner Y matches only in science-category subjects, and is thus assigned 0.5 points. Consequently, at this stage, Learner X will be a learner with similar attributes. The available attributes continue to be added up in this way. Ultimately, learners with the highest scores are extracted as learners with similar attributes.

Table 1: Attribute table

	Science-category class	Natural Science	Physics	Chemistry	Mathematics
N	0.5	1	2	0	0
X	0.5	1	2	0	0
Y	0.5	0	0	0	1

4.5 Relationship between Similar Learners and Users with Similar Attributes

The flowchart in Figure 3 shows that when a sufficient learning history is available, the attribute correlation method is bypassed and recommendations are based on the normal algorithm for collaborative learning recommendations. This is because it can be surmised that the attribute correlation method will not produce better results by extracting similar learners based on learning history. This is because the recommended content itself is used as the history when extracting similar learners. In comparison, the background information of the learner, which has no direct relation, is used with attribute correlation. When these two approaches are compared, the learning history clearly constitutes pure information in terms of the system. For example, in order to recommend books to a person who has not read any books to date, the thinking is incorporated that books will be recommended that may appeal to that person's preferences, based on elements such as other interests and skills. The primary aim of this method is to solve the Cold-Start problem.

5 TESTING

Testing was conducted on subjects to clarify the outcomes of the proposed method. The following two items were evaluated.

- Is the proposed method effective?
- Was the hypothesis pertaining to attribute data selection proven?

5.1 Hypothesis Pertaining to the Selection of Attribute Data

As described in Section 3.1, attribute data serve as the meta-data for learners. However, not all of the personal data of learners is necessarily required. For example, if one were recommending exercises to help a person stay fit, physical information such as height and weight would be important, but this type of information is not necessary when recommending novels. In other words, it was theorized that attributes that are relevant to the content being recommended will probably demonstrate a high correlation. Here, because we are creating a recommendation system to be used in an education support system, information relating to learning will demonstrate a high correlation compared to attributes that are not particularly related to learning.

5.2 Test Content

Advance preparation: To prepare for testing, courses from a high school curriculum were systematized as attributes related to learning, and hobbies were systematized as attributes other than learning-related attributes. The reason for choosing hobbies as attributes was that learners acquire and actively choose hobbies, as opposed to inherent information such as height, so these were assumed to closely reflect learner preferences. High school courses were selected as learning attributes in order to eliminate differences based on school year, since the students taking part in the testing were university students. As no models existed that were systematized with respect to hobbies,

systematization was done based on speculation. For high school courses, however, we referred to the “Senior High School Education Guidelines” issued by the Ministry of Education, Culture, Sports, Science and Technology, as shown in [8]. Attribute hierarchies were each organized into three layers, with the objective of suppressing any bias created by differences in scores occurring as a result of changes in the weight of scores based on the depth of the hierarchy layer. Attribute data obtained as a result consisted of two attributes and three hierarchical layers.

Subjects: Subjects were grouped into two groups comprising a total of 18 students, and a questionnaire was conducted prior to the testing. Participants answered the following two questions.

- What were your strong subjects when you were in high school? (Learning attributes)
- What are your current hobbies? (Hobby attributes)

Attributes of subjects were compiled based on the questionnaire. As a large number of attributes could be selected, the questionnaire was conducted in a self-reporting format, but in cases where the student did not respond correctly, that student was asked the question again by the tester, for the purpose of normalizing the attribute information. Subsequently, the following three items pertaining to the contest of the test were explained to the subjects, and testing was conducted.

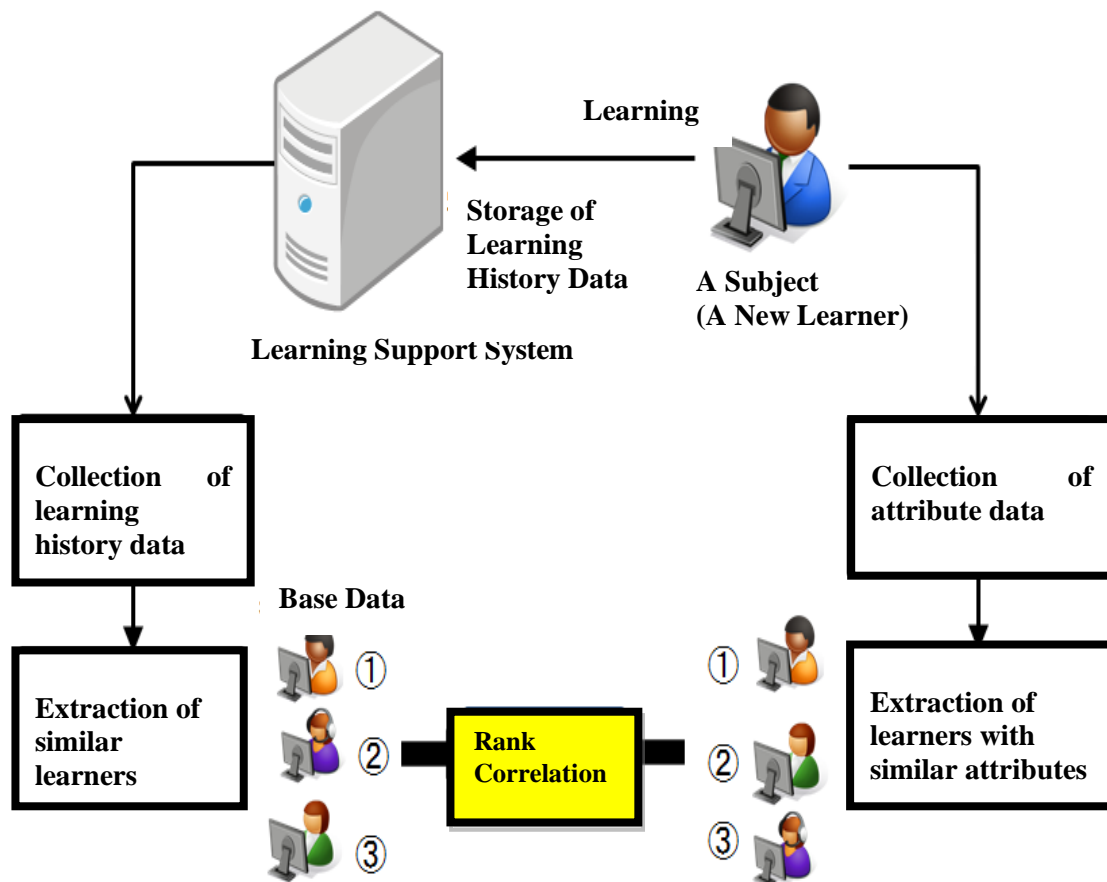


Figure 3: Outline of experiment.

- Learning time would be 15 minutes.
- Content would be in the form of a database.
- An achievement test would be performed after the study time had ended.

Moreover, the database comprising the content was something that could not be learned in its totality in 15 min, so subjects were asked to select portions that they did not understand, and to focus on those items when learning. This was done in order to avoid having subjects start at the beginning and study the contents in sequential order. The achievement test was also designed to increase the motivation of subjects to study efficiently in a short period of time, and would not affect the test evaluation itself.

Analysis method: Figure 3 shows a schematic for the testing. The degree of similarity (similar learners) was calculated based on the learning history obtained from the 15-min period of learning, and the degree of similarity (learners with similar attributes) was calculated based on the compiled attribute data. The two were then compared and evaluated. Specifically, the same number of rank correlations was acquired as the number of subjects, and correlations were acquired in relation to the rankings of similar learners and learners with similar attributes obtained from each of the two similarity scales noted above. The Jaccard coefficient was used to calculate the degree of similarity based on learning history, as described in Segaran[9], and Kendall's rank coefficient correlation was used to calculate the rank correlation, as shown in Iwasa and Yadohisa [10]. The attribute correlation method is designed only to address new learners. The degree of similarity based on learning history shows a high degree of reliability, and so was used as the reference. In other words, the aim was to obtain the rank correlation between the ranking for the degree of similarity based on learning history (similar learners) and the degree of similarity calculated based on the proposed method (learners with similar attributes), so if the average of all subjects was high, reliability in terms of the extraction of similar learners would be seen as high, and the approach could be considered effective.

5.3 Test Results

Table 2 and Table 3 show test results for the two groups. The figures represent mean and standard deviation for the group as a whole, calculated based on the rank correlation between the ranks of learners with similar attributes and those of learners with similar learning histories. As the rank correlation is a correlation coefficient, values were taken from between -1 and 1. The closer the value is to 1, the stronger the correlation. The closer the value is to -1, the stronger the inverse correlation. The closer the value is to 0, the weaker the correlation. As can be seen from the two tables, the average was |0.1| or less for both, so no correlation was demonstrated, and no significant results were obtained. Moreover, with respect to learning attributes and hobby attributes, the only differences were due to error, so the hypothesis was negated. Except for one item, standard deviations were all ≤ 0.2 as well, indicating that this conclusion is appropriate.

Table 2: Experimental results for group 1

	Learning Attribute	Hobby Attribute	Whole Attributes
Average	-0.077	0.044	-0.095

Table 3: Experimental results for group 2

	Learning Attribute	Hobby Attribute	Whole Attributes
Average	0.1032	0.0238	-0.0397

5.4 Discussion

Considering the causes of the results produced, the possibility arises that the amount of attribute data was insufficient. In that light, looking at the individual data for each subject, in the rankings based on attribute correlation, it was seen that rankings at the same ratio occurred for many subjects. Among these, there were a number of cases in which hobby attributes ended up being the same numeric values as those for other subjects as a whole, and no ranking correlations could be determined. However, despite the small volume of sample data, the fact that the average value for correlation coefficients was close to zero cannot be ignored. One other problem was that the relationship between the content being recommended and the attribute data was not clear. As indicated in Section 4.5, the reliability of attribute data is unclear, from an objective standpoint.

6 COLLABORATIVE ATTRIBUTE METHOD

In the collaborative attribute method of testing described in Section 5, usefulness could not be confirmed, for the reasons described in Section 5.4. Given that, we used the background information as attribute data. The collaborative attribute method is proposed here as a method for extracting new learners.

6.1 Overview

Using the background information of the learner as attribute data is the same approach used in the attribute correlation method. The attribute correlation method consisted of systematizing this data before use, but the data are not systematized in the method proposed here, but rather used as is. The degree of similarity between learners is surmised with reference to the degree of similarity between learners based on learning history, and to the attribute data.

6.2 Degree of Similarity between Attribute Data

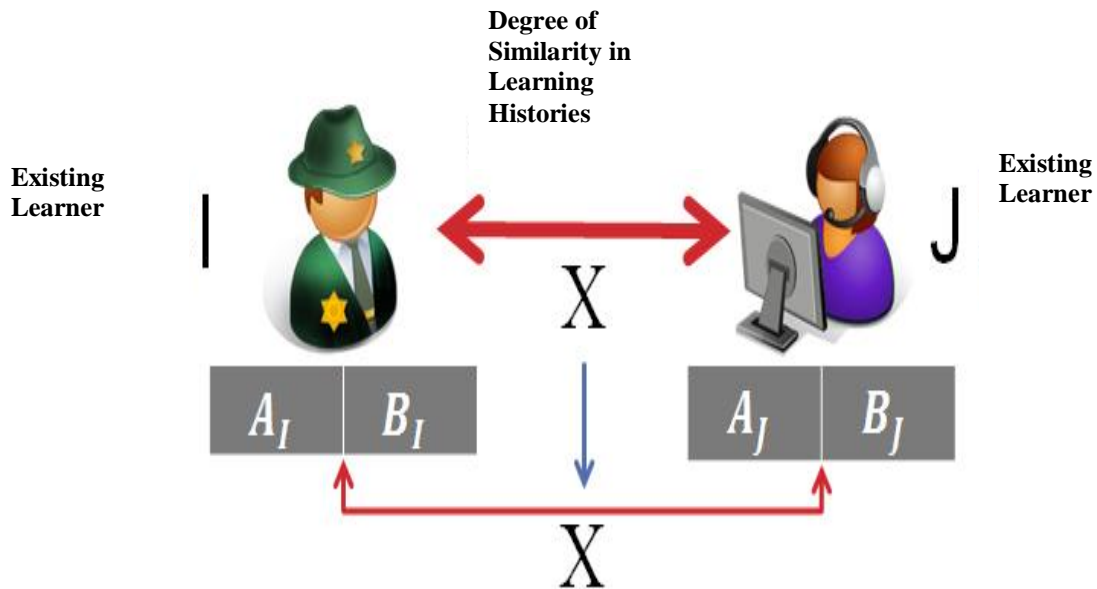


Figure 4: Calculating similarities between attribute data.

The degree of similarity between attribute data was calculated in advance. Here, we take Learner I and Learner J, for whom a certain amount of learning history has been compiled. Attribute data for these two users were acquired when they were new learners, so we already have degrees of similarity in learning histories and respective attribute data at this stage. Amounts of attribute data are not determined in particular, but let us assume in this example that we have two attribute data: A and B. Taking the degree of similarity in learning histories between these two persons as X, we can say that the combination of attributes for these two persons, for some reason, has similarity X. If this combination is also seen among other learners, we take the average. These degrees of similarity are then accumulated in a database. Figure 4 shows a schematic diagram of this.

When actually making recommendations for new learners for whom no degree of learning history similarity exists, we refer to similarities between attribute data that have been accumulated, and extract learners having combinations with the highest degrees of similarity between attribute data as learners with similar attributes, as shown in Figure 5. Content is then recommended based on these users.

6.4 Differences between This Method and the Attribute Correlation Method

In the attribute correlation method, attributes are systematized and the number of points is totaled. In the collaborative attribute method, however, similarities between attributes are measured using similarities in learning histories, which are reliable, as a resource. As a

6.3 Deriving the Degree of Similarity

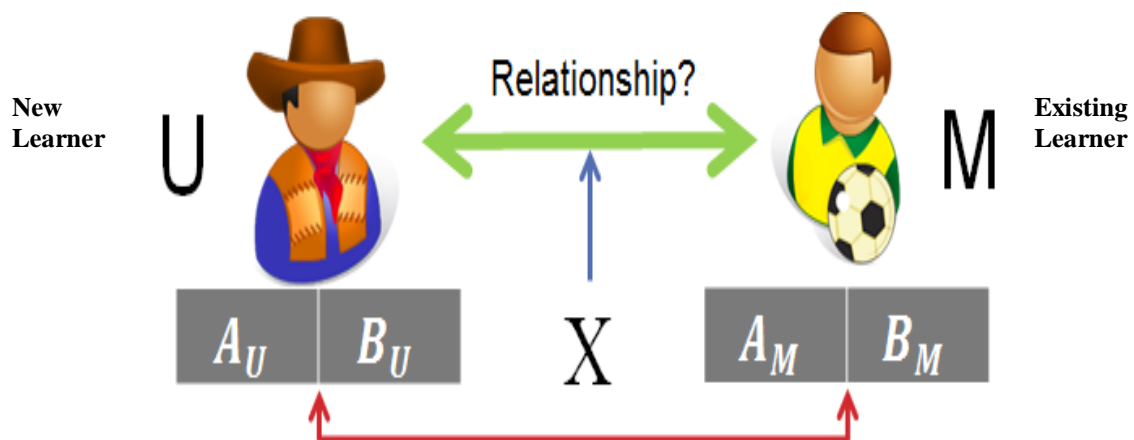


Figure 5: Extract the attribute analogy.

result, the data can be expected to be more reliable. Conversely, because the approach taken is similar to that in CF, recommendations will similarly be less accurate if only small amounts of data have been accumulated.

6.5 If the Amount of Attribute Data Accumulated Is Insufficient

As indicated in Section 6.4, this method also involves accumulated attribute data, and there are concerns that the extraction of persons with similar attributes will be less accurate if insufficient information is available. If the amount of attribute data accumulated by forming combinations of attributes of a learner for whom recommendations are being made is smaller than a stipulated amount, attributes A and B are split and calculated, as shown in Figure 6.

Now, assume that we want to find the similarity of A_x and A_1 . We load combinations that include and from a table of attribute data similarities that have been accumulated, and we take the similarity of each of these and divide the number of points by the ratio of the number of elements. For example, if the ratio of the number of elements of A and the number of elements of B is 1:2, and the similarity between $A_x B_a$ and $A_1 B_n$ is 0.6, this result of 0.6 would be divided by 1/3 to obtain a result of 0.2. This would be carried out for the number of combinations A_x of and A_1 , and the average of all values would be taken. This would be done as many times as there are combinations of the attributes of A and B, and recommendation content would be extracted from users having the combinations with the highest values.

6.6 Experimental Results

We experimented with the collaborative attribute method using the same data as those in Section 5. Table 4 shows the results of rank correlations.

Table 4: Experimental results of rank correlation with collaborative attribute method

	Rank Correlation
Average	0.188
Average	0.237

We can see that the collaborative attribute method provides better results than the attribute correlation method in Table 4, but they are not so high values. Some values of the rank correlations which are above 0.4 exist among the results before averaging. So, we can expect the averaged rank correlation will be higher if we can collect more data.

7 CONCLUSION

In the testing described in Section 5, the usefulness of the attribute correlation method was not able to be proved. This was attributed to the fact that the relationship between attribute data and learning history is not understood, and a collaborative attribute method is proposed in which similarities in learning history are referenced and attributes are used. At the same time, however, this method has not yet been perfected and still has scope for improvement. In addition, it may simply be that not enough testing has been conducted on the attribute correlation method. In the future, we intend to continue conducting testing on the attribute correlation method, and to develop, implement and test a collaborative attribute method.

8 ACKNOWLEDGMENT

This research was supported by one Grant-in-Aid for Scientific Research C (Subject No. 21500908), and is currently supported by the other Grant-in-Aid for Scientific Research C (Subject No. 24500122).

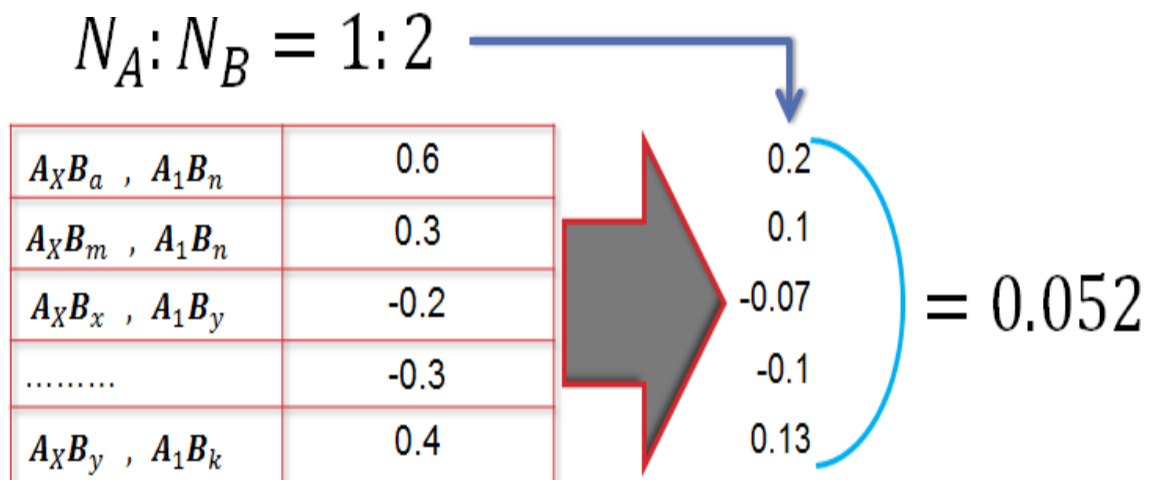


Figure 6: Algorithm for split attributes.

REFERENCES

- [1] Y. Wada, S. Matsuzawa, M. Yamaguchi, and S. Dohi, Bidirectional Recommendation Technology for Web Digital Texts, *Journal of Digital Information Management*, vol. 8, no. 4, pp. 240–246, August 2010.
- [2] Y. Wada, Y. Hamadume, S. Dohi, and J. Sawamoto, Technology for Recommending Optimum Learning Texts Based on Data Mining of Learning Historical Data, *International Journal of Information Society IJIS.*, vol. 2, no.3, pp.78–87, April 2011.
- [3] Y. Wada, T. Segawa, J. Sawamoto, and H. Sato, RESEARCH ON RECOMMENDER FUNCTIONS FOR LEARNING SUPPORT SYSTEMS, 2nd International Conference on Applied and Theoretical Information Systems Research, December 27-29, 2012.
- [4] Amazon: <http://www.amazon.com>
- [5] P. Resnick, N. Iacovou, M. Suchak, P. Bergstorm, and J. Riedl, GroupLens: An Open Architecture for Collaborative Filtering of Netnews, *Proc. ACM Conf. on computer Suspend Coparative Work*, pp.175-186, 1994.
- [6] S. Kitamura, Consistency Between Theoretical Interests in Collaborative Learning Studies and Methods of Statistical Analysis: A Review of Statistics for Hierarchical Data, *Japan Society for Educational Technology (ISSN 1349-8290)*, vol.33 , no.3 , pp.342-352, 2010.
- [7] A. Schein, A. Popescul, L. Unger, and D. Pennock, Methods and Metrics for Cold-Start Recommendations, *25th Annual ACM SIGIR Conf.*, pp253-260, 2002.
- [8] http://www.mext.go.jp/b_menu/shuppan/sonota/990301d.htm
- [9] T. Segaran, *Programming Collective Intelligence Building Smart Web 2.0 Applications*, O'REILLY (ISBN 978-4-87311-364-7), 2008-7.
- [10] H.Iwasa and H.Yadohisa, ISBN 978-4-7980-2396-0, p.352, 2009-10.

Application of a Lump-sum Update Method to Distributed Database

Tsukasa Kudo[†], Yui Takeda[‡], Masahiko Ishino^{*}, Kenji Saotome^{**}, and Nobuhiro Kataoka^{***}

[†]Faculty of Comprehensive Informatics, Shizuoka Institute of Science and Technology, Japan

[‡]Mitsubishi Electric Information Systems Corporation, Japan

^{*} Department of Management Information Science, Fukui University of Technology, Japan

^{**} Hosei Business School of Innovation Management, Japan

^{***} Interprise Laboratory, Japan

kudo@cs.sist.ac.jp

Abstract - At the present time, with spread of the internet business, many business systems have become to be built as distributed systems. Accordingly, the database is also dispersed to plural business systems as the distributed database. By the way, in the actual business systems, a lump-sum update of a great deal of data has often to be performed concurrently with the online transactions. However, in this case, there is a problem of efficiency in the conventional update methods, which update databases by the chain of many divided transactions. For this problem, we have shown the efficient lump-sum update method for centralized business systems, which use the transaction time. However, as shown by the distributed transactions, some transaction features for the distributed database are different from the centralized database. Therefore, in this paper, first we show the problems on applying this method to the distributed databases. Secondly, we propose their measures. Moreover, through the evaluations using a prototype, we confirmed that our measures are valid and this method can be applied to the distributed databases.

Keywords: database, distributed database, distributed transaction, business system, nonstop service.

1 INTRODUCTION

With spread of the internet business and decentralized technology, many business systems have become to be built as distributed systems at the present time. That is, each system performs its business by mutual cooperation with other systems [7], [11]. For example, in corporations, business systems are built at each branch office, and the business of this office is performed using its system. On the other hand, each business system is operated as a part of the distributed system, since each system accesses the data of the other systems when it is necessary. In this way, the distributed business system as the whole corporation is composed. And, it is possible to reduce the communication cost and perform the suitable system operation for the each office. On the one hand, as for the online services on a wide range of Internet such as net shops, non-stop services have become common because of the convenience of customers, the globalization and so on. So, it is difficult to stop the online service for the particular business.

However, in such a business system, a great deal of data often has to be updated in a lump-sum. So, formerly, it was executed as a night batch to avoid the time zone of the online service. However, because of the spread of nonstop service,

it has to be executed concurrently with the online service at present. Therefore, some methods were put to practical use to execute it concurrently with the user entry of online service (hereinafter “online entry”). Here, these conventional methods divide the updating of a great deal of data into short time transactions, and then execute and commit them one after another. Therefore, though its impact on the online entry is small, there are problems: the intermediate results of the updating can be queried by the other transactions, that is, the isolation cannot be maintained; its processing efficiency declines because of the increase of its commit number.

For these problems, we had proposed an updating method that utilizes the records about the transaction time[2]. Moreover, we had shown the following evaluation result as for the updating of a great deal of data in the centralized systems: it executes the update more efficiently than the conventional method with maintaining the ACID property[3]. Since the transaction time is a kind of time of the temporal database, hereinafter we call it “temporal update” method. However, in order to apply this method to a distributed database, it is necessary to measure against not only the distributed transaction but also the various problems of the distributed environment.

Here, the temporal update has the characteristic of its process as follows: the completion time of updating is set beforehand; as for the commit of this method, its execution control has to be performed for the serialization between the online entries. However, there are problems: the dispersion in the update time is often increased by the efficiency of the network and cooperative business system environments; the synchronization between plural servers often causes the decline of efficiency. Therefore, in this paper, we propose a method for applying the temporal update method to the distributed databases. Moreover, we show the following experimental results of the prototype: the problems can be solved by the proposal method; since the implementation of the destination server of data transfer is easy, this method is valid for a data distribution system having many destination servers.

The remainder of this paper is organized as follows. Section 2 shows the related works about the update transaction; an overview of the temporal update method; the problem about applying it to the distributed databases. In Section 3, we propose the method for this problem, and show its implementation in Section 4. In Section 5, we show the experimental results of the prototype, and show our considerations in Section 6.

2 PROBLEM OF DATABASE UPDATE IN A LUMP-SUM

2.1 Related Works

As for the update of the database, it is necessary that the transaction maintains the ACID properties: atomicity, consistency, isolation, and durability[1]. Here, since a lot of users use the online entry concurrently, a lot of corresponding transactions access the database concurrently. Therefore, database updates are serialized by locking the data to be updated by each transaction, and we can obtain a result as if transactions were executed sequentially. On the other hand, in the actual business systems, it is necessary to update a great deal of data in a lump-sum. For example, in the banking systems, the ATM is provided as the online service, and a lot of users perform online entry at the same time. On the other hand, a great deal of account transfer that is entrusted from the credit card company and so on is executed as the lump-sum update.

Since the update time of such a lump-sum is often so long, there is the problem that it can't lock the whole target data to serialize between the online entries. It makes the online entries wait for a long while. So, some methods were put to practical use to execute it concurrently with the online entries. In the mini-batch, a great deal of data is divided into small units, and they are updated and committed individually to shorten each update time. That is, the lump-sum update is performed by a set of short transactions. Also, in sagas, they compose a sequence of transactions and are performed one after another. It has a configuration to recover by executing the compensating transaction corresponding to each transaction in the case of fault [1], [10]. But, in these methods, the update and commit are repeated alternately many times. And, it makes the problems: the intermediate results of the updating can be queried by the other transactions; the efficiency declines because of the increase of commit number.

Here, in the distributed databases, the transaction needs not only the update and commit feature as for the individual database but also the feature for simultaneous update of multiple databases. So, the distributed transaction feature was put to practical use, in which the concurrency control across multiple databases is performed by two-phase commit and so on [6]. In this way, as for the distributed database, the different transaction feature from the centralized database has to be introduced.

By the way, the temporal database was proposed from the viewpoint of the record management about the time [8]. The transaction time is one of the times of this database: the time that a fact is valid in the database, which is expressed by the half-open interval $[T_a, T_d)$. Here, T_a shows the addition time that the data of the fact was added to the database; T_d shows the deletion time that it was deleted from the database. Even in the case of data deletion, data is deleted only logically by setting the deletion time, so the data record is left. Incidentally, if the data wasn't deleted, the value of attribute T_d is expressed by *now* [9]. It shows the current time and changes with passage of time. The relation R_t of the table having the transaction time is expressed below.

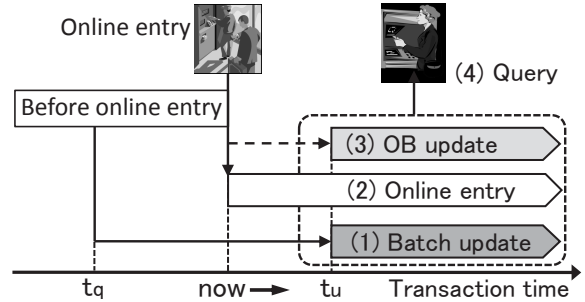


Figure 1: Change of data by temporal update method.

$$R_t(K, T_a, T_d, A) \quad (1)$$

Here, T_a and T_d are the above-mentioned transaction time. Let $q[T_a]$ be the value of attribute T_a of tuple q . Then the data set of the snapshot of R_t at the designated time t is expressed by the following $Q(t)$.

$$Q(t) = \{q|q \in R_t; q[T_a] \leq t \wedge t < q[T_d]\} \quad (2)$$

Here, K is the primary key attribute of the snapshot; A is the other attribute. The transaction time is not exposed to the users. Since the update of the online entry is performed at the current time $t = \text{now}$, the data of Equation (2) at any past time can be queried without the conflict with the online entry.

For a data update method, we can use the optimistic concurrency control [4] by utilizing the transaction time. The lump-sum updates are often performed by the following procedure: reading the target data; generating the update data from it; updating the database. In this method, the transaction confirms the transaction time of the target data again at the update timing. And, if it was not changed from the read timing, it shows the data was not updated by another transaction. However, the transaction needs to lock the data between this confirmation and the commit.

For another update method which utilizing a time, there is the timestamp-ordering concurrency control. It uses the time stamps (start time) of transactions $\{T_1, T_2, \dots, T_n\}$, which compose the ordered set. The time stamp is stored in data when a transaction accesses the data, and the order of transactions that access to each data is maintained by it [4]. However, when one transaction is updating a data, the other transactions that update the same data have to wait until the commit.

Thus, these methods intend for short time transactions. For example, since the lump-sum update takes a long while, the other transactions are also waited for a long while. That is, it disturbs the non-stop services. Moreover, we can't find the lump-sum update method for a great deal of data with maintaining the ACID property.

2.2 Temporal Update Method

As for the centralized database, we proposed the temporal update method that utilizes the transaction time to update a great deal of data in a lump-sum with maintaining the ACID property of the transaction [2]. Figure 1 shows the data change with the transaction time in the case of the database

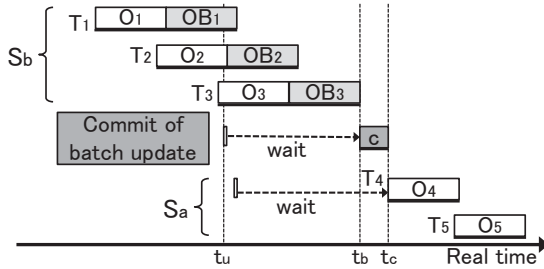


Figure 2: Serialization of batch update and online entry.

update by this method. In this method, the lump-sum update is executed by both the batch update (1) and OB update (the online batch update) (3). Here, the batch update (1) corresponds to the usual lump-sum update. It updates the data at the past time t_q , and stores the update results at the future time t_u that was “beforehand” established. Incidentally, though the online entry (2) is executed even during the batch update (1), it updates the data at the current time *now*. So, the competition between (1) and (2) can be avoided. On the other hand, since the batch update also has to update the data changed by the online entry, the OB update performs the process corresponding to it individually.

As a result, at time t_u , three kinds of results are stored: the batch update, the online entry and the OB update. Therefore, we query only the high-priority data by the following order of priority, and we can get the query result as if the batch update and online entry were performed in a series.

- (A) First, for each value of the primary key K of Equation (1), we select the data that was updated at the latest time. Here, we use t_q to the updated time as for only the batch update.
- (B) Second, if there are plural data having the same key value, we select the data by the following priority of update process: the OB update, the online entry, the batch update.

In the case of Figure 1, since both results of the OB update and online entry were updated at the latest time, the OB update result is queried based on above-mentioned (B). Also, if there is no online entry as for the case of Figure 1, there are the online entry data entered before t_q (“Before online entry” in the Figure), and the result of the batch update. So, the batch update result is queried based on above-mentioned (A). That is, in this method, all the updated data are stored, and only the valid data is queried. As a result, we can perform the lump-sum update and the online entry concurrently, without their competition.

However, at the end of a batch update, the control for the serialization between it and the online entry is necessary. That is, the online entry, which is begun before time t_u , is accompanied by the OB update to reflect the batch update. On the other hand, another one begun after t_u has to be performed using the result of the batch update. We show this in Figure 2. Therefore, in this method, the commit of the batch update is executed after the online entries (S_b) that are begun before time t_u ; the other online entries (S_a), which are begun after t_u , are waited until this commit.

2.3 Problem about Application to Distributed Database

As for the temporal update method, we had been assuming the centralized database and the controlled operation. That is, as mentioned in Section 2.2, we assumed that the batch update time t_u can be established beforehand; the individual online entry completes in a short time. In the actual systems, the former feature is often used in the lump-sum update executing at the designated time: the bank transfer at the designated time; the change of organization data at the designated date; and so on. However, as for the case where we apply this method to the distributed database, it has to be executed in the various environments. So, the following problems occur.

First, the dispersion of the batch update time is very larger than the centralized database. Though the batch update updates plural databases at the same time, the individual environment in this distributed system is varied: the traffic on the network; the load and performance of each server. So, the prediction of the time t_u is extremely difficult. Therefore, in the case that the prediction time is earlier than the actual elapsed time, the batch update doesn’t complete by t_u . So, it is aborted, and it has to be re-run. On the contrary, if it is later, the unnecessary OB update must be continued after the batch update completion. As a result, there is a problem that the efficiency as the whole system declines.

Second, the problem that the online entry wait of one server spreads to the other servers occurs. As for the distributed database, the related online entries executing in all the servers have to wait the completion of the commit of the batch update until t_c as shown in Figure 2. However, since the systems were built individually at each branch office, there may be the online entries having a long time transaction. Therefore, there is a problem about the delay of the commit of the batch update, which is caused by some online entry. And, it delays the online entries in the all related servers.

3 PROPOSAL METHOD FOR DISTRIBUTED DATABASE

To solve the problem for applying the temporal update to the distributed database, we propose the following two methods.

3.1 Setting Method of Dynamic Batch Update Completion Time

For the problem about the elapsed time of the batch update, we propose the method to perform its commit immediately after the batch update. Here, as shown in “Table” of Figure 3, since the predicted completion time is set to every updated data by the batch update and OB update, it takes time to update these time again. Therefore, in this method, we use a view table to change these times when these data are queried.

Figure 3 shows the overview of this method. Here, time is shown as date: year, month and day. The predicted completion time of the batch update is set as the temporary time, and its first digit is replaced by “@” as an example. In the case

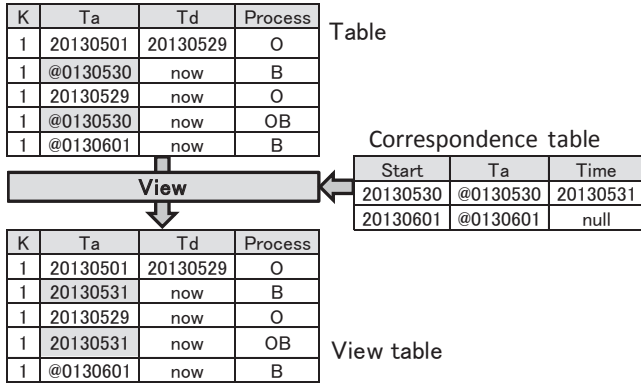


Figure 3: Change of addition time of business data.

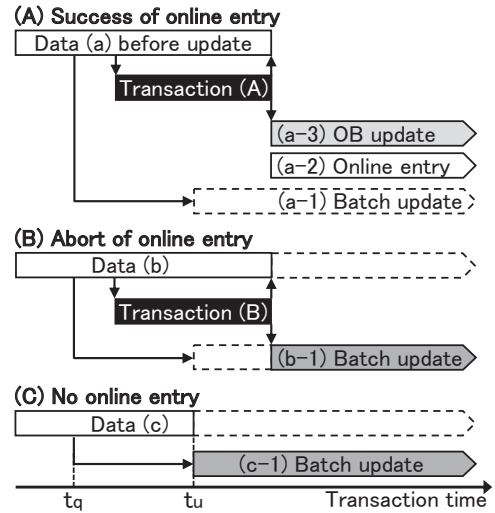
of Figure 3, the addition time T_a , which is equal to the completion time, is set to “@0130530” and “@0130601”. Also, its “Process” column shows the classification of the update process of the data: O (the online entry), B (the batch update) and OB (the OB update). In this case, the data entered by the online entry on 5/1 of 2013 was updated by the batch update on 5/29, which above-mentioned completion time (T_a) is “@0130530”. And, the OB update was performed along with this. In addition, the batch update with the completion time “@0130601” was performed after this.

When the batch update completed, its completion time is set (“Time” of “Correspondence table”), which corresponds to the temporary time T_a . Incidentally, it is set to *null* until this completion. Only if the completion time is set to “Time”, T_a of data of “View table” is replaced with it. In the case of Figure 3, “@0130530” of this is replaced with “20130531”; “@0130601” isn’t replaced. That is, even if there are many target data, all T_a of them can be replaced by updating only one data. Therefore, this update process can be executed by a short time transaction. Incidentally, since the value of “@” is larger than any numerical value, the data having the temporary time (with “@”) isn’t queried by Equation (2) with designating time.

3.2 Serialization Method without Lock Feature

To solve the problem shown in Figure 2, we propose the serialization method between the online entry and batch update. In (A) and (B) of the Figure 4, we show the case that the online entry is executed across the batch update completion time t_u . Incidentally, (C) shows the case that there is no online entry, as the reference. Here, the black hatching shows a transaction of the online entry; the broken line shows the data that should not be queried though it is exists.

First, (A) of Figure 4 shows the success case of the online entry. Transaction (A) continues across the batch update completion time t_u . If the batch update result ($a-1$) is queried between time t_u and the completion time of this transaction, it becomes a phantom read. Therefore, the mechanism, which makes this data not to be queried, is necessary to maintain the consistency. Incidentally, after the transaction completed, data (a) is deleted logically by setting the deletion

Figure 4: Query result as for online entry spanning t_u .

time; the online entry data ($a-2$) is inserted; the OB update data ($a-3$) is inserted if data (a) was the target of the batch update. So, ($a-1$) is not queried, though either ($a-2$) or ($a-3$) is queried.

Next, (B) of Figure 4 shows the abort case of the online entry. Transaction (B) continues across the batch update completion time t_u similar to (A), and its result is undecided at t_u : success or abort. Therefore, as for data ($b-1$), the mechanism similar to (A) is necessary. In the case of the abort, data (b) remains without change, since the rollback of the transaction is performed. And, since the batch update was executed, its result ($b-1$) has to be queried after the completion of the transaction. That is, for the serialization, the temporal update has to be composed to obtain the following query result.

- **During the online entry:** the batch update result is not queried, but the data just before the start of the online entry transaction is queried.
- **After the online entry completion (success):** either the online entry result or the OB update result is queried. Incidentally, in only the case of this data being target of the batch update, the latter occurs.
- **After the online entry completion (abort):** the batch update result is queried.

Therefore, in the case that the online entry transaction continues across the batch update completion time t_u , the consistency of the query result is maintained by the mechanism: the batch update result is not queried until the completion of the transaction. Therefore, we propose a method using a table that manages the key of the data being updated by online entry transactions. And the data, which key is registered in this table, is excluded from the query result of the batch update. This table has the following relation.

$$R_e(t_name, t_key) \quad (3)$$

Here, t_name shows the name of the target table; t_key shows the value of the primary key of the online entry data.

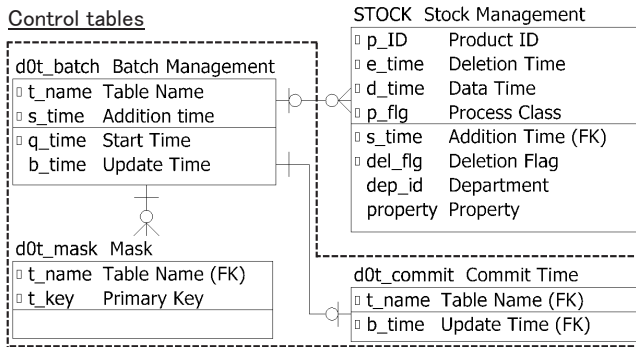


Figure 5: ER diagram of control tables.

And, (t_name, t_key) composes the primary key of this table. In this way, it is possible to perform the serialization by the table R_e without the lock feature between the onlien entry and batch update.

4 IMPLEMENTATION

First, Figure 5 shows the ER diagram of the Control tables, which are used for the implementation of the proposal method. These tables save the data to control the temporal update and are placed in each target database of the temporal update. Batch management table $d0t_batch$ manages the time of the temporal update, and it saves the following data: Table name t_name ; Addition time s_time , Start time q_time ; Update time b_time . Here, b_time is used to set the completion time of the batch update at the timing of its completion dynamically as shown in Figure 3. That is, if the estimation of the completion time of the batch update is difficult beforehand, a temporary time and null are set to each s_time and b_time at the timing of its beginning. Then, the completion time t_u is set to b_time at the timing of its completion to query the updated data having the addition time t_u . Also, Mask table $d0t_mask$ is an implementation of the relation of Equation (3). Commit time table $d0t_commit$ stores the latest completion time t_u in b_time for each target table which name is set to t_name . And the result of the batch update and OB update which s_time is before b_time is queried.

Stock Management table $stock$ is an example of the business data table, and we use it for the experiment in Section 5. Here, as for the stock management, various kinds of distributed databases are used. For example, in the retail companies, each branch has its database system and manages its stock. On the other hand, the database serves as a part of the distributed database in the case of the stock movement among branches or the delivery from the distribution sector. Similarly, in the manufacturing industry, the supply chains of the parts are built among the related companies.

As for the attributes of $stock$, the following are correspond to the attributes (K, T_a, T_d, A) of the relation R_t shown in Equation (1): Product ID p_ID ; Addition Time s_time ; Deletion Time e_time ; Department dep_id and Property $property$. In addition, we add the following properties for the temporal update. Data Time d_time shows the update order of the data as shown in Section 2.2; Process Class p_flg show the classi-

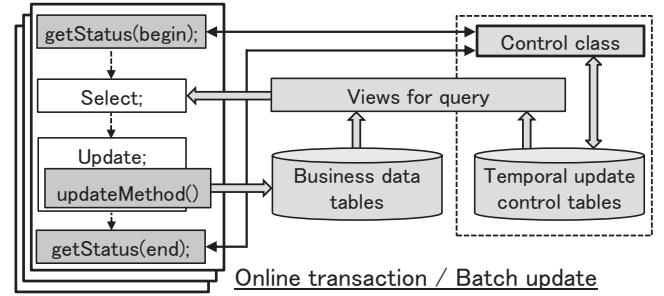


Figure 6: Composition of temporal update.

fication of the update process of the data shown as “Process” column in Figure 3; Deletion Flag del_flg shows the data was deleted if it is set. Here, del_flg is used to exclude all data having the same primary key K from the query result. For example, in the case that the OB update (3) is deletion in Figure 1, the other data (1) and (2) also not have to be queried. For this purpose, we exclude the unnecessary queried data (3) after the query, and as a result, the other data is also not queried.

Second, the concurrency control between the online entries and batch update is necessary as shown in Figure 4. For this purpose, we implemented the Control class by the Java to perform this control as shown in Figure 6. It exposes a method $getStatus$ that is the interface of business programs, and the programs call this method at the timing of the start and completion each of the online entry transaction and batch update. Then, the updating of the control tables and the control about the serialization are performed by this method.

In the case of the batch update, $getStatus(begin)$ is called at the timing of start to set the control data of the batch update to $d0t_batch$ of all related databases. $getStatus(end)$ is called at the timing of its completion similarly, and it sets the completion time to b_time of $d0t_batch$ and updates $d0t_commit$. Thus, the batch update result can be queried. In the case of the online entry, $getStatus(begin)$ is called at the timing of its transaction start similarly, and confirms whether or not the batch update updates the target table. If it is not being updated, only the usual online entry transaction is performed. On the contrary, if it is being updated, the transaction’s data is set to $d0t_mask$ by this method. And, the transaction performs the OB update after the online entry, and calls $getStatus(end)$ to delete the data of $d0t_mask$ at the timing of its completion.

Third, The business programs query the table through its view table, if it is the target of the temporal update. The view table is created by the DDL (Data Definition Language) shown in Figure 7 and has the following feature.

- It changes the addition time T_a of the view table as shown in Figure 3.
- It controls query results of the batch update and OB update by using the time changed in (a): only the data, which Addition Time s_time is older than Update Time b_time of $d0t_commit$ (including b_time), is queried. Incidentally, every online entry result is queried.
- It excludes the data from the query result of (b) by using

```

CREATE VIEW stock_v AS
SELECT p_id, COALESCE(b.time, a.s.time, b.s.time),
       e.time, d.time, p_flg, del_flg, dep_id, property FROM stock a
LEFT OUTER JOIN d0t_batch b USING (s.time)
WHERE p_flg= '1'
OR (p_flg = '0' AND p_id NOT IN
    (SELECT t_key FROM d0t_mask
     WHERE t_name = 'stock' AND q.time = a.d.time)
OR p_flg = '2')
AND COALESCE (b.time, a.s.time, b.s.time) <=
    (SELECT b.time FROM d0t_commit
     WHERE t_name = 'stock');
    
```

Figure 7: Example of DDL to create view table.

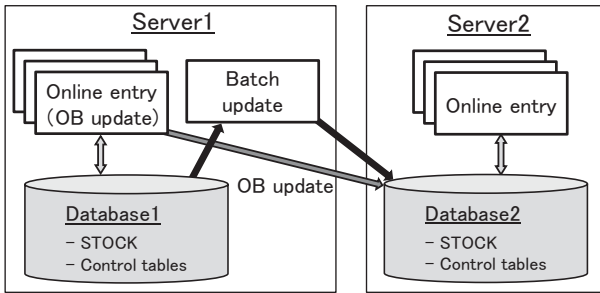


Figure 8: Experimental system composition.

d0t_mask, which is being updated by the online entry transaction.

Using these view table, there are the following effects: since it doesn't allow to query the intermediate state of the temporal update shown in Figure 4, the ACID properties of the database can be maintained; the developer can build the program efficiently by it, since it conceals the above-mentioned procedure.

On the other hand, these view tables aren't always updatable. So, though it is necessary that the table is updated directly by the business programs, transaction time attributes of the table are not exposed to users: *e.time*, *d.time* and *s.time* of *STOCK*. Therefore the method *updateMethod* is provided, and the business programs update tables using this method. In addition, the above-mentioned features about databases were implemented by MySQL: InnoDB for the transaction feature; XA transaction for the distributed transaction [5].

5 EXPERIMENTS AND EVALUATIONS

5.1 Overview of Experiments

To evaluate the proposal method, we built the prototype of a stock management system, and performed experiments by it. We show the composition of the experimental system in Figure 8. This system is a distributed system consisting of two servers, and each server has its own database. Its business table is the stock management table shown in Figure 5. And, we assume this system is a non-stop service system. That is, its data can be deleted by the online entry transactions at any time when the corresponding goods are sold.

Data	Time (Sec)								Remarks
	2	4	6	8	10	12	14	16	
← Batch update →									
Server1 (Local)									
R010	●	●	●						Sold
R011	●	●	●	●	●				Sale canceled, Move out
R020	●	●	●	●	●	●			Sold
R021	●	●	●	●	●	●			Sale canceled, Move out
R100	●	●	●	●	●				Move out
Server2 (Remote)									
R010									
R011					●	●	●		Move in
R020									
R021						●	●		Move in
R100					●	●	●		Move in

Figure 9: Query result of proposed temporal update.

The stock movements from *Database1* to *Database2* are performed in a lump-sum when necessary. This is built by using the temporal update method, and its batch update results, which is shown by (1) in Figure 1, are inserted to *stock* of each database. Among them, as for the data of *Database1*, Deletion Flag *del_flg* is set not to be queried after the completion of the temporal update. On the one hand, the data is inserted into the table of *Database2*, and they can be queried after the completion time. In addition, in the case that the data of *Database1* is deleted by the online entry, its movement to *Database2* has to be also canceled by the OB update. Concretely, as for *Database2*, the data with setting *del_flg* is inserted by the OB update, so the batch update results are also excluded from the query result.

In the experiment, we performed the online entry in both of the server, and performed the above-mentioned temporal update on its way. In Figure 9, we show the query result of the typical data along the passage of time. Among the data, *R010* was sold during the data movement by the temporal update; the sale of *R011* was canceled, that is, the transaction was canceled by the rollback. As for *R020* and *R021*, the sale of them continued across the completion time of the data movement, and *R020* was sold; the sale of *R021* was canceled. *R100* was only moved. Also, the black circle shows that the data was queried at the time through the view table in Figure 7; the blank shows not to be queried;

5.2 Evaluations of Validity of Proposal Method

First, in order to complete the temporal update immediately after the batch update, we set the temporary time “@0” (we show only its second, the same in the following) at the timing of the batch update start. Then, we set its completion time “11” to Update Time *b.time* of both of *d0t_batch* and *d0t_commit* at the timing of the batch update completion. As shown by *R011* and *R100*, we could complete the temporal update immediately after the batch update and query their update result.

Next, for the serialization between the online entry and batch update without the lock feature, we inserted *d0t_mask* the corresponding data to the online entry at the timing of its

start; we deleted the data at its completion. As a result, during online entry, the batch update results are not queried, but the data before the online entry is queried as shown by *R020* and *R021*. That is, as for the sold data *R020*, the data before its selling is queried until the completion of the online entry; it can't be queried after the completion. Also, it didn't be moved to *Database2*. As for *R021* in the case of sale being canceled, since it was moved at the timing of the completion of the online entry transaction, it can't be queried in *Database1* after the completion; it can be queried in *Database2*. As described above, the temporal update results are queried after the online entry transaction, that is, the serialization between them could be controlled.

5.3 Evaluations of Implementation in Distributed Environment

As for the data movement by the temporal update, its business programs in the destination server could be composed by only local transactions. That is, as shown in Figure 8, as for Stock Management Table *stock* in *Database2*, since the existing data in another database is only inserted by both of the batch update and OB update, there is no competition with the online entry. Also, since Control tables are used only inside of the view table and hidden from the above-mentioned programs, these tables could be implemented without influences on the existing programs of *Server2* except the implementation about the view tables.

On the other hand, as for the business program in the server where the temporal update is performed, the implementations about it were necessary. As for the updating of Control tables, by developing the control class for the temporal update, the business programs could be configured to call its methods at the timing of start and end of the online entry transactions as shown in Figure 6. So, these tables are hidden from the business programs. However, since the OB update has to be executed as a part of the online entry transaction, it had to be integrated into the corresponding online entry program. By the way, as for the implementation of this method, since the data of *d0t_mask* has to be inserted before the online entry transaction start, the transaction of this method had to be executed separating from it. Similarly, in the case of abort, since the rollback of the online entry transaction is executed, the data of *d0t_mask* had to be deleted by another transaction. In addition, in the case of success, the deletion could be executed in the same transaction.

As for the batch update, the serialization control between the temporal update and the online entry transactions had to be executed at the start and end of the transaction. On the other hand, since this control wasn't necessary except these timing, the business tables of each database could be updated by the local transactions one after another. In particular, in the case of the number of the updating data was small, its commit was not necessary in the middle of the updating. So, the updating features could be implemented by only using SQL statement without the cursor operations, and we could implement it more efficiently than the mini-batch.

6 CONSIDERATIONS

We found that the temporal update can be completed immediately after the completion of the batch update by this method. As for the temporal update, since the OB update has to be executed as a part of the online entry transaction, its execution time is considered to become long particularly in the distributed environment. Therefore, from the viewpoint of the efficiency, to reduce the execution time of the temporal update is effective.

Incidentally, we consider that the method to reserve the completion time of the temporal update is also useful in both of the centralized and distributed systems. It can be used in the case to change great deal of data in a lump-sum at the designated time and so on. For example, we discuss the case that the share of product of each branch is changed at the prearranged date in Figure 8. In this case, we execute the temporal update with reserving its completion time at 0 a.m. of the designated date and the stock of each branch office can be updated at once with reflecting the sale earlier.

Also, we found that the serialization between the temporal update and online entry can be composed without the lock feature. It has been pointed out that the long time transactions and the lock feature cause the fault in the distributed systems. So, we consider that the lump-sum update can be composed more secure by this method.

Moreover, in the case of the data movement from one server to another server, we found that we need to implement only the control tables and the view tables as for the latter; the existing business programs are not affected except the implementation about the view table. In particular, even the lock feature for update isn't necessary, because the lump-sum update can be executed by only the insertion of data. That is, this method is valid in the case of data transfer from the specified administration server to the other servers widely.

7 CONCLUSIONS

In IWIN2012, We showed the temporal update method in a centralized database is effective in the viewpoint of maintaining the consistency and updating data efficiently. However, to apply this method to the distributed database, there are problems: it is difficult to estimate its completion time; an online entry wait is spread to the other servers at the completion timing of this method, because the serialization between the batch update and the online entry transactions has to be performed. For these problems, we propose the method in this paper for the following purpose: the beforehand estimation of its completion time becomes unnecessary; the serialization can be executed without the online entry waits. And, we confirmed by experiments that these feature was valid and could be implemented in the distributed database. Moreover, we find through these experiments this method is also valid in the case of data transfer from the specified administration server to the other servers in a wide area.

The future challenge is the evaluation of the operational efficiency and performance in the viewpoint of the business system in order to adopt it to the actual distributed systems.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 24500132.

REFERENCES

- [1] Gray, J., Reuter, A., Transaction Processing: Concept and Techniques, San Francisco: Morgan Kaufmann (1992).
- [2] Kudo, T., et al., A batch Update Method of Database for Mass Data during Online Entry, Procs. 16th Int. Conf. on Knowledge-Based and Intelligent Information & Engineering Systems – KES 2012, pp. 1807–1816 (2012).
- [3] Kudo, T., et al., Evaluation of Lump-sum Update Methods for Nonstop Service System, Procs. Int. Workshop on Informatics (IWIN2012), pp. 3–10 (2012).
- [4] Lewis, P. M., Bernstein, Arthur., Kifer, M., Databases and Transaction Processing: An Application-Oriented Approach, Addison-Wesley (2001).
- [5] ORACLE, XA Transactions, <http://dev.mysql.com/doc/refman/5.1/en/xa.html>.
- [6] Özsu, M. T., Valduriez, P., Principles of Distributed Database Systems, Springer (2011).
- [7] Shanker, U., Misra, M., Sarje, A.K., Distributed real time database systems: background and literature review, Distributed and Parallel Databases, Vol. 23, Issue 2, pp. 127–149 (2008).
- [8] Snodgrass, R., Ahn, I., Temporal Databases, IEEE COMPUTER, Vol. 19, No. 9, pp. 35–42 (1986).
- [9] Stantic, B., Thornton, J., Sattar, A., A Novel Approach to Model NOW in Temporal Databases, Procs. 10th Int. Symposium on Temporal Representation and Reasoning and Fourth Int. Conf. on Temporal Logic, pp. 174–180 (2003).
- [10] Wang, T., Vonk, J., Kratz, B., Grefen, P., A survey on the history of transaction management: from flat to grid transactions, Distributed and Parallel Databases, Vol. 23, Issue 3, pp. 235–270 (2008).
- [11] Yang, J., Lee, I., Jeong, O., Song, S., Lee, C., Lee, S., An architecture for supporting batch query and online service in Very Large Database systems, IEEE Int. Conf. on e-Business Engineering (ICEBE '06), pp. 549 – 553 (2006).

Comparative analysis of cognition and memorization during learning using simple electroencephalographs

Koji Yoshida¹, Fumiyasu Hirai¹, and Isao Miyaji²

¹ Shonan Institute of Technology

1-1-25, Tsujido Nishikaigan, Fujisawa, Kanagawa 251-8511, Japan

² Okayama University of Science, 1-1 Ridaicho, Okayama, 700-0005, Japan
yoshidak@info.shonan-it.ac.jp, fumi_3070@yahoo.co.jp, miyaji@mis.ous.ac.jp

Abstract—Simple electroencephalographs (EEG devices), which have recently been used commercially to an increasing extent, are portable and wearable, such that they do not restrict a wearer's actions. This convenience of use allows the ordinary use of electroencephalography inexpensively and widely. This study examines the construction of a system that can feed back EEG information obtained using a simple EEG devices for instruction assistance in distance learning. This paper presents a discussion of the characteristics of a simple EEG device and the state of EEG, and describes an experimental comparative analysis of correlation of cognition and memorization during a student's learning act with EEG data obtained by EEG measurement. The results revealed the effectiveness of β/α as an index for stress and attention level in cognition. Results reveal that Low- γ , which reflects memorization work, is effective as an index for estimating memorization-oriented learning approaches.

Keywords: Brain wave sensor; Meditation, α wave and β wave, Distance Learning, e-Learning

1 INTRODUCTION

Electroencephalography (EEG), which provides biological information, is widely used as a performance index of information processing taking place in the brain. Frequency response among EEG characteristics is known to be related closely to cognitive processes such as learning, language, and perception [1]. By virtue of continued development of brain science and technology, electroencephalographs (EEG devices) that used to be expensive and bulky have been miniaturized for portability. Simple EEG devices that are wearable and sufficiently compact to permit a wearer to have unrestricted movement have become commercially available recently. We specifically assess the merits of a simple EEG device and explain the construction of a system that feeds back EEG information for use with distance learning.

Distance learning systems are beneficial because the progress and results of learning can be fed back and checked immediately. However, there are shortcomings: student learning cannot be observed directly, and insufficient information such as the learning state and progress

information can offer only limited support. It is therefore indispensable to observe the cognition and mental condition of user students with biological information obtained from EEG, to enable the support of students in light of their actual conditions. Such a system for observation can be expected to improve distance learning shortcomings and to encourage instruction assistance and student learning.

This paper presents a discussion of the characteristics of a simple EEG device and the state of EEG, describes an investigation of whether EEG information is useful as an external index of cognition state at EEG measurement, and explains an experimental comparative analysis of correlation of cognition and memorization during a student's learning with EEG data obtained from EEG measurements.

2 ELECTROENCEPHALOGRAPHY

Electroencephalography (EEG), which can measure an index of human information-processing steps, is widely employed in medicine for integrative functional evaluation of a brain, and for investigation of brain disorders through epileptology and angiopathy. An electrical signal arises in the event of neural ignition or synaptic transmission in a brain. This biological signal can be recorded using an electrode placed on the scalp as it emerges as brain potential change. It is designated as an electroencephalogram [2]. The EEG data are classified into five types according to the frequency range. Listed below are the designation, frequency range, and typical mental state of the appearance of each wave.

- δ waves, 1–4 Hz, in sleep
- θ waves, 4–8 Hz, in sleep/attention
- α waves, 8–12 Hz, relaxed/eyes closed
- β waves, 15–20 Hz, concentrated/moving
- γ waves, 30 Hz -, processing memory and vision

Fourier analysis of original collected EEG data yields the power spectrum of each frequency.

However, EEGs show many individual differences. The relation between EEG and the cognition state varies with circumstances even for the same person. The emergence of α waves does not always imply that a subject is relaxed. Accordingly, it is necessary to perform repeated measurements and to compare data with EEG obtained in various circumstances.

3 SIMPLE ELECTROENCEPHALOGRAPH

An EEG device measures and records electroencephalographic data. Conventional experimental studies of brain physiology use a large-scale apparatus. However, such an EEG device is unsuitable for ordinary use. Medical-grade instruments with many electrodes bother subjects because of their inconvenient requirements for wearing and restriction of movement, which burden subjects with stress. This situation might inhibit their learning, which is the objective of this study. When medical level precision is necessary for acquisition of EEG data, EEG equipment should be used. However, small portable EEG devices that are easily available are desirable in cases where EEG information is used for applications, assuming a simple EEG input interface and ordinary use.

For these reasons, it is easy and effective to use a simple EEG device rather than the medical type of EEG device for introducing EEG devices to educational use, as in this study. Therefore, this study conducts EEG experiments using MindSet™ produced by NeuroSky, Inc. [3], which is inexpensive and wearable. MindSet transmits digital EEG data to a PC. The potential between a sensor on the forehead and an electrode on the ear is measured, collected EEG data are analyzed with an on-board chip built in an ear pad, and data are transmitted to a PC using Bluetooth, a wireless communication system. Figure 1 shows communication with MindSet and a PC.

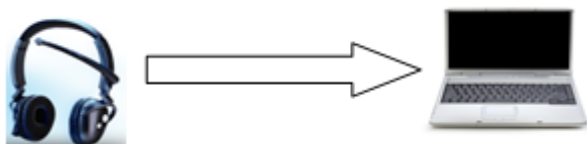


Figure 1: MindSet™ and communication.

Features of MindSet are listed below.

- Measurement point: at the frontal lobe with one sensor (international 10 / 20 system (Fp1)).
- A reference point is set on an earlobe.
- Dry sensor type EEG module.
- A chip in the ear pad performs from sensing to analysis.
- Operable with most processors and DPS.
- Data transfer to PC employs Bluetooth communication.
- Sampling at 512 Hz.
- Each frequency component is extracted using fast Fourier transform (FFT) for every second.

A sampling frequency of 512 Hz assures 512 original EEG data obtained every second. Frequency components are extracted by application of FFT to these data, which are then digitized and transmitted to a PC. Other signals are transmitted and received as data aside from these, including poor-sig-lev (noise intensity) and e-sense meter (an original index of NeuroSky) such as the attention level and meditation level. Table 1 shows the range of each frequency component at FFT.

Table 1: Frequency component table of brain wave

Type	Measurable data (Hz)	State of mind
δ waves	0.5–2.75	Deep sleep without dreaming, non-REM sleep, unconsciousness
θ waves	3.5–6.75	Intuition, creativity, remembrance, imagination, illusion, dreams
Low α waves	7.5–9.25	Relaxed but not lazy, peace, consciousness
High α waves	10–11.75	Formerly designated as sensorimotor rhythm (SMR), relaxed but concentrating, integrative
Low β waves	13–16.75	Contemplation, recognizing self and environment
High β waves	18–29.75	Alert, wakefulness
Low γ waves	31–39.75	Memorization, high-order cognitive activity
Mid γ waves	41–49.75	Visual information processing

There are libraries and applications accompanying the simple NeuroSky EEG device, which can facilitate users' research and development. The system environment of this study collects EEG data using an application provided by NeuroSky.

4 STATE OF EEG AND LEARNING STATE

Previous research findings in psychology and brain science empirically teach that EEG waveforms are useful as an index of a mental condition if observed with a related event. The measurement of the following is regarded as effective to observe human mental conditions: The power spectrum of α and β waves obtained by discrete Fourier transform of obtained EEG, the fraction of α or β waves to the whole EEG, and the ratio of α waves to β waves [4,5]. α waves are a waveform that is generally observed during rest and wakefulness. However, the α wave amplitude is generally enlarged in a relaxed state, but it shrinks with tension and the appearance of β waves.

Particularly β waves (13–30 Hz) are considered to be closely related to cognition states. Some reports describe studies that address the relation between intellectual tasks and EEG. Giannitrapani et al. [6] measured the EEG of healthy persons during an intelligence test. They discovered that the low-frequency component of β waves became predominant during reading tests, mathematics tests, and a figure alignment tests, but they are less dominant during other tests, which demonstrates that β waves are effective to some extent as an index for inferring a cognition state.

It is assumed that γ waves (31 Hz or over) are related mainly to higher order mental activity, which is regarded as a mechanism that yields the variance matrix of cognitive processing. It allows synchronous and concerted cognitive activity such as perception. Because γ waves are prone to be affected by motion of muscles and eyeballs, it is important that signals be separated carefully for identification.

5 EEG ACQUISITION SYSTEM

It was necessary to construct an EEG acquisition system to record EEG in advance of EEG observation experiments in this study. Because the API of MindSet can acquire but not record EEG data, a Windows program that records EEG data was coded in Java and implemented. Figure 2 depicts the schematic diagram of the system.

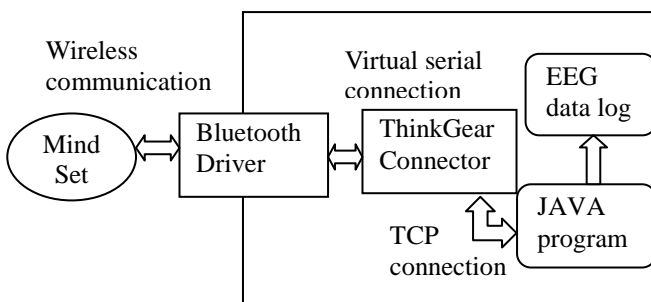


Figure 2: Schematic diagram of EEG Acquisition system.

MindSet transmits data to a PC through a Bluetooth Driver. A server program distributed by NeuroSky (Think Gear Connector) is accessed by TCP when extracting data from the PC. The Java program prepared in this study performs socket communication to the server program and receives data. A packet is transmitted every other second and the acquired packet is analyzed. This provides the numerical data, attention level, relaxation level, and sensor sensitivity of each frequency range. Because the received data are big endian, log data are written as a text file in little endian, floating point format for processing by this system.

6 EEG MEASUREMENT EXPERIMENT

This experiment is aimed at analyzing the relation between an EEG and cognition state in a learning state under cognitive work with a simple wearable EEG device.¹ Accordingly, an EEG under cognitive work is measured, the relation between the EEG measured, cognition and frequency response are observed, and the correlation between a learning state and an EEG is investigated.

EEG measurement was made possible by cooperation of several students in our laboratory as subjects. Measurements were conducted in a seminar room, and attention of the subjects was controlled by providing sufficient intermissions.

Power spectra at respective frequency ranges and sensor sensitivities were recorded during experiments. Analysis was conducted only when sensor sensitivity was the best, as a precondition of analysis. δ and θ wave regions below 5 Hz were excluded from analysis because they are prone to be affected by noise. The mid- γ wave region (41–49.75) was also excluded because it is affected strongly by muscle and eyeball motion. The Low- γ wave region (31–39.75) was analyzed because it is close to the β wave region and is influenced slightly. Otherwise, strange numerical values occurred among continuous and stable data even at the best sensor sensitivity, although rarely. Such cases were excluded as noise.

6.1 Experiment Outline

- Subject: Three men in their 20s (university natural science students)
- Measuring time: Up to 3 min
- Cognitive themes:
 - (Experiment 1) Music appreciation
 - (Experiment 2) Learning of arithmetic themes (flash mental calculation)
- Supplement: Survey after experiments

6.2 Experimental Procedure

A subject who is seated on a chair in front of a PC executes a specified cognitive task wearing the simple EEG device. The electroencephalogram is produced by the EEG device. Then, the standard deviation of each frequency parameter and the intensity of each power spectrum are analyzed comparatively using spreadsheet software. Because the taste of genres in music appreciation and the calculation approach in free calculation might depend on each subject, ambiguous parts were supplemented by survey from a subject after experiment.

- Taste of genres from three kinds of music appreciation
- Approach at free calculation
- Good or poor at calculation and memorization

6.3 Experiment 1 – Music Appreciation

Subjects listen to music in designated genres for 3 min. Measurement is conducted with eyes open. The genres to be appreciated are as follows.

- Music appreciation (electronic)
- Music appreciation (classical)
- Music appreciation (jazz)

(1) Experimental result

Table 2 presents results of music appreciation based on the survey result after experiments. The table has a ranking of each subject's favorite genre, the spectral average of α and β components, and the average of β/α . Figure 3 is a graph showing the ranking of each subject's favorite genre and the average of β/α .

Table 2: Result of music appreciation.

	Music genre	Favorite ranking	α Average	β Average	β/α Average
Subject A	Electronic	1	3.35895	1.74865	0.65167
	Classical	3	2.36037	1.39479	0.78461
	Jazz	2	3.25979	1.77580	0.67513
Subject B	Electronic	3	2.21645	1.32405	0.82492
	Classical	1	2.67938	1.56875	0.72554
	Jazz	2	2.39551	1.45120	0.79272
Subject C	Electronic	2	2.69348	1.59780	0.76520
	Classical	3	2.38589	1.40642	0.80344
	Jazz	1	2.82039	1.61367	0.74251

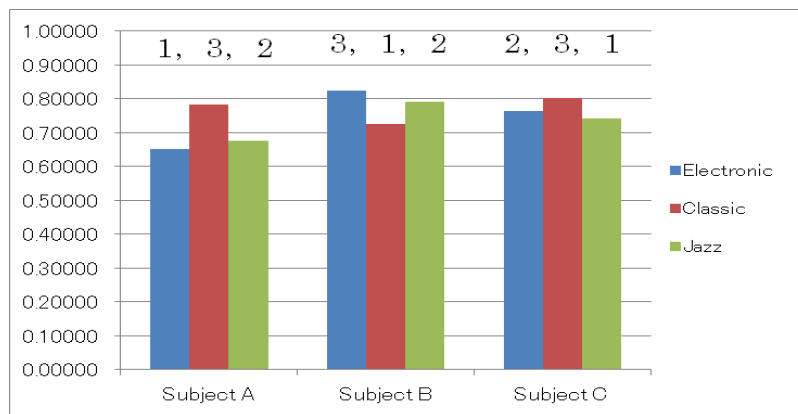


Figure 3: Favorite Genre ranking and β/α average.

6.4 Experiment 2 – Arithmetic Theme

Subjects must solve 10 problems, each requiring addition of five numbers continuously displayed at an interval of 0.5 s. The problems are given one task after another with an interval of 3 s, conducted under the following guidance:

- Free calculation (no guidance provided, each subject can tackle the problem freely)
- Calculation-oriented (compute tasks one by one as soon as they are displayed without memorization)
- Memorizing-oriented (compute tasks after memorizing five numbers displayed in advance)

(1) Experimental result

The approaches the subjects used to solve free calculation problems were assessed. Most subjects selected the calculation-oriented or memorization-oriented approach according to a general situation to tackle problems, in spite of differences in detail.

Subject A: calculation-oriented; memorizes and then computes only when pushed for time.

Subject B: calculate first up to the second – third numbers, then memorizes remainder and computes.

Subject C: calculates only numbers that are easy to compute, and if judged as difficult, then memorizes and calculates.

It was also surveyed how good or poor the subjects were at calculation and memorization. Tables 3-1, 3-2, and 3-3 show experimentally obtained results for this point. The table has β/α and the average and dispersion of Low- γ of the following three subjects considering results of Experiment 1: Subject A, who favors electronic music and who is good at calculation; Subject B, who favors classical music and who is good at memorization; and Subject C, who favors jazz music and has no special capabilities for calculation and memorization. Figures 4 and 5 respectively show comparisons of β/α and the average of Low- γ of each subject.

Table 3-1: Subject A, who favors electronic music and is good at calculation.

Subject A	Cognitive work	β/α Average	β/α Dispersion	Low- γ Average	Low- γ Dispersion
Music appreciation	Electronic	0.65167	0.23039	0.28196	0.05579
	Classical	0.78461	0.25888	0.30471	0.09386
	Jazz	0.67513	0.20000	0.22446	0.04573
Arithmetic theme	Free calculation	0.70166	0.22966	0.24043	0.07013
	Calculation-oriented	1.57748	1.74876	0.27906	0.04940
	Memorization-oriented	0.86462	0.22815	0.77958	0.69460

Table 3-2: Subject B, who favors classical music and is good at memorization.

Subject B	Cognitive work	β/α Average	β/α Dispersion	Low- γ Average	Low- γ Dispersion
Music appreciation	Electronic	0.82492	0.26036	0.25213	0.02883
	Classical	0.72554	0.27589	0.24604	0.03976
	Jazz	0.79272	0.21514	0.21514	0.06640
Arithmetic theme	Free calculation	0.92881	0.74894	0.53062	0.39876
	Calculation-oriented	1.09457	1.18760	0.40270	0.14323
	Memorization-oriented	0.74728	0.24607	0.80152	0.74628

Table 3-3: Subject C, who favors jazz music and is both so-so at calculation and memorization.

Subject C	Cognitive work	β/α Average	β/α Dispersion	Low- γ Average	Low- γ Dispersion
Music appreciation	Electronic	0.76704	0.28385	0.30986	0.05646
	Classical	0.80344	0.33070	0.28253	0.05584
	Jazz	0.74251	0.31665	0.30328	0.04828
Arithmetic theme	Free calculation	1.18927	0.99259	0.37377	0.09034
	Calculation-oriented	1.37995	1.62096	0.40329	0.08138
	Memorization-oriented	0.84637	0.37325	0.62736	0.59020

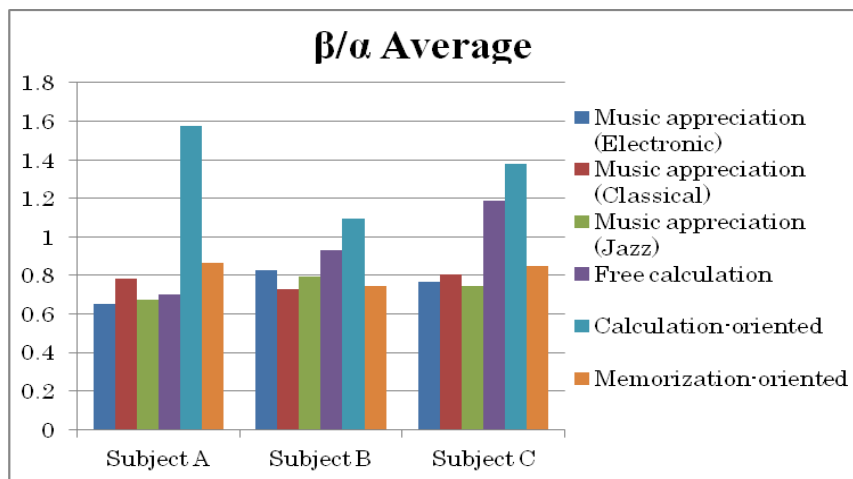


Figure 4: Average of β/α .

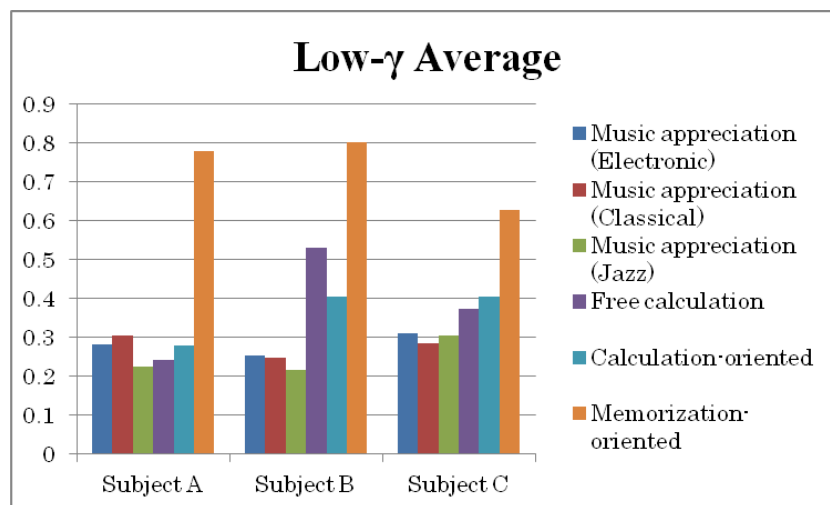


Figure 5: Average of Low- γ .

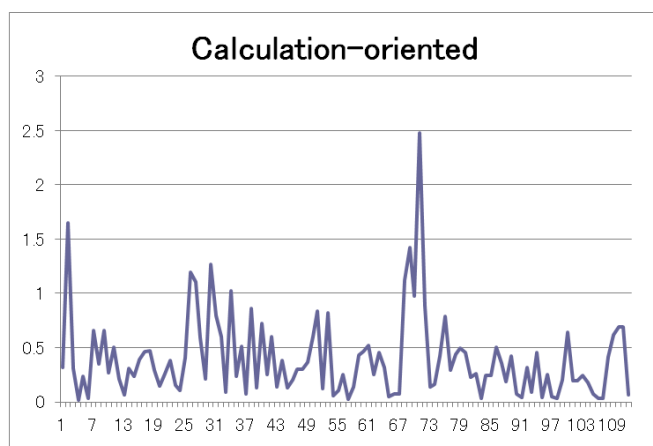
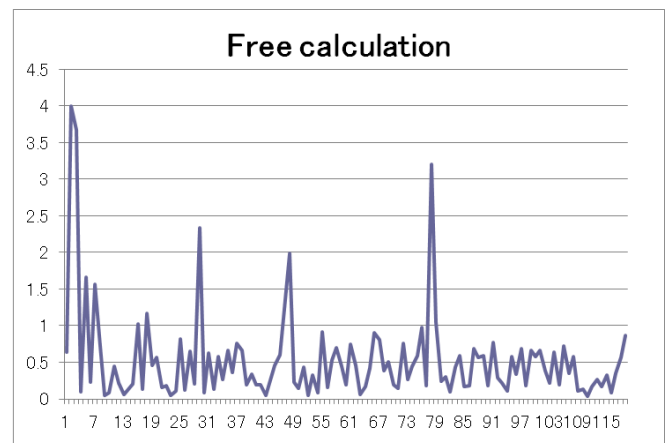
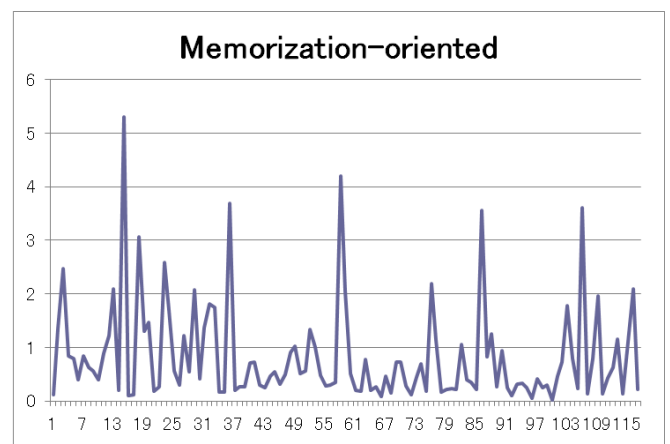
(2) Analysis of experimentally obtained results

The experimentally obtained results show that β/α of the arithmetic theme was considerably greater than that of music appreciation. Table 3 and Figure 4 demonstrate a significant upward tendency such that the average of β/α of all three subjects exceeded 1 in the calculation-oriented approach.

The attention level of β/α varied with how good or poor a subject was at the task. Change in subject A, good at calculation, was the greatest; then C and B followed, which is the same order as goodness in calculation. The change of β/α in the memorization-oriented approach was about as much as that in music appreciation, which revealed that β/α varies greatly without any modifying difficulty in an experiment theme depending on an approach.

Next, the Low- γ component is discussed. Table 3 and Figure 5 demonstrate an increase in Low- γ component instead of change in β/α in the memorization-oriented approach. The average and dispersion of Low- γ component increased from calculation-oriented, free calculation, to memorization-oriented. Especially, the dispersion of the Low- γ component in memorization-oriented changed greatly.

Subject B, who is good at memorization, provided interesting results. Figures 6-1, 6-2, and 6-3 show the transition of Low- γ component in the arithmetic theme of Subject B. Low- γ component was observed to change greatly in the descending order of calculation-oriented, free calculation, and memorization-oriented. The maximum of Low- γ component in memorization-oriented approach was more than twice that of calculation-oriented approach. Furthermore, the graph in the memorization-oriented approach had such a drastic change that the Low- γ component increased whenever a new problem was given.

Figure 6-1: Low- γ of Subject B (calculation-oriented).Figure 6-2: Low- γ of Subject B (free calculation).Figure 6-3: Low- γ of Subject B (memorization-oriented).

7 COMPREHENSIVE EVALUATION AND DISCUSSION

(1) Music appreciation

Comparison of genre taste, α waves and β/α of each genre in music appreciation revealed that β/α was high for a non-favorite genre, but low for a favorite genre, in spite of little change in magnitude. It is considered to be true because music of the favorite genre was pleasant and sufficiently comfortable that β/α decreased. Consequently, this result suggests that β/α is effective in estimating stress level representing pleasant and unpleasant sensations depending on individual tastes.

(2) Arithmetic theme

β/α tended upward because the arithmetic theme required cognition compared with music appreciation. Comparison of calculation-oriented, free calculation, and memorization-oriented indicates significant increase in β/α in calculation-oriented that requires cognition and concentration most, which suggests that β/α increases when cognition and concentration are needed to activate the brain.

Comparison of subjects by whether good-or-poor in calculation shows that the variation of β/α was influenced thereby. In the calculation-oriented approach, β/α was large when a subject was good at calculation, although it was small when a subject was poor at it. However, β/α changed

little in the memorization-oriented approach, which is assumed because memorization freed a subject from instantaneous calculation and became unhurried. Consequently, the change was slight.

Presumably, β/α changes greatly when instantaneous cognition (flash mental calculation) is required and the brain is concentrated, whereas change in β/α decreases when a subject is not pressed, considering the above and unmodified difficulty of the experiment theme. Accordingly, β/α is regarded as effective as an index that measures the attention level for instantaneous cognition.

It was also revealed that the more that memorization was required, the stronger the Low- γ component was, in the order of memorization-oriented, free calculation, and calculation-oriented, which indicates that a Low- γ component responds in the event of work to memorize. Consequently, a Low- γ component is presumed to be effective to measure the degree of the memorization-oriented approach. Furthermore, instantaneous memorization required in the experiment is assumed to correspond to the short-term memory in cognitive psychology. Accordingly, it is necessary to consider experiments on the transition by the amount and difficulty of the content to be memorized in the future.

8 CONCLUSION

The comparative analysis of the power spectrum of α and β waves and β/α in multiple cognitive processes was conducted experimentally in this paper, for the correlation analysis of contemplation in learning state using a simple EEG device. It is presumed that observation of the β/α component is effective as an index to measure the degree of stress and attention level for instantaneous cognition. In addition, the Low- γ component is presumably effective as a criterion of judgment for the measurement of the degree of memorization-oriented approach.

An experiment equipped with the EEG device, an unfamiliar device, did not readily comfort subjects to their respective ordinary states. For that purpose, it was presumed necessary to let subjects get used to the simple EEG device first by increasing the occasions of measurement. Detailed analysis of these results is expected to elicit a still more effective index to assess the attention level of a brain and memorization.

Consequently, detailed analysis and application study of these devices and methods will be conducted in the future, aimed at employing them for feedback information used with a distance-learning system. This study was supported by a grant-in-aid for scientific research No. 24501219.

REFERENCES

[1] Y. Sakamoto, K. Yoshida, and I. Miyachi, comparative Analysis of Thinking at Learning State by the Simple Electroencephalograph, Multimedia, Distributed, Cooperative, and Mobile Symposium (DICOMO) 2012, Symposium Proceedings, pp.724- 729 (2012).(in Japanese)

[2]K. Kitajo, and Y. Yamaguchi, Study on Visual Perception by EEG phase synchronization analysis, Vision Vol. 19, No. 4, pp.193-200 (2007). (in Japanese)

[3] NeuroSky, Inc. <http://www.neurosky.com/>

[4] Uwano, Ishida, Matsuda, Fukushima, Nakamichi, Ohira, Matsumoto and Okada: "Evaluation of Software Usability Using Electroencephalogram – Comparison of Frequency Component between Different Software Versions," Human Interface Society, Vol. 10, No. 2, pp. 233–242 (2008). (in Japanese)

[5]K.Yoshida, Y.Sakamoto, I.Miyaji, K. Yamada: "Analysis comparison between α wave and β wave at the learning status by simple electroencephalography", KES'2012, Proceedings, Knowledge-Based Intelligent Information and Engineering Systems, pp.1817-1826 (2012).

[6] D. Giannitrapani, The role of 13-hz activity in mentation, The EEG of Mental Activities, pp. 149-152 (1988).

Sugoroku Game Interactions with Twitter

Jun Munemori*, Kanako Morimoto*, Junko Ito*

*Faculty of Systems Engineering, Wakayama University, Japan
{munemori, s145050, itou}@sys.wakayama-u.ac.jp

Abstract – We propose a web service that incorporates Twitter with sugoroku, which is a traditional Japanese game. Twitter poses problem questions, and players of the game answer it. The problem questions are contributed regularly. An avatar in the sugoroku game moves forward when a player answers a question correctly. When two or more players are at the same position, he/she can view information about the other player(s) and contact them via Twitter. We applied this service twice in an experiment, and the results indicated that users found the game interesting (4.6/5.0) and that the interaction between Twitter and the sugoroku was rated highly (3.8/5.0).

Keywords: Twitter, quiz, interaction, sugoroku

1 INTRODUCTION

Services that provide a way to communicate through the Internet have dramatically increased their users in the past decade [1]. Twitter [2] is one of those services, and clients who can read a tweet are distributed throughout the world. (Twitter is a registered trademark of Twitter, Inc.). In addition, many services have been developed to allow users to collaborate via Twitter, including games. Some services post the result of the game as a tweet. However, only a few services exist where Twitter is an integral part of the game. That is, most services using Twitter are one-way services. Twitter and its interactive service are unknown. Two-way (or interactive) services between Twitter and other services are requested.

Therefore, we propose a service with a strongly linked interaction between Twitter and the sugoroku, which is a traditional Japanese grid game.

2 RELATED WORK

2.1 Twitter

Twitter is a service administered by Twitter, Inc., where the user can publicly post a “tweet,” which is a short sentence that is 140 characters or less. Users can share a tweet with other users as a post to announce or initiate a conversation, or they can use a tweet to reply to another. The Twitter API is provided to third-party applications to be able to access Twitter functionality.

2.2 Dotwar

Dotwar [3] is a game that uses a user’s profile image from Twitter. The profile image of the user is divided into dots, which become a group of soldiers, who fight against the dots from the profile image of another user. The profile image is acquired automatically when a user enters a Twitter account ID in the game. The user can also tweet the result of the play. This game is not an interactive game completely.

2.3 Furai no shiren 4+

Furai no shiren 4+ [4] is an RPG game. Users search in dungeon, which is generated at random. The game system itself is not related to Twitter, but the result of the game could be tweeted by the user. That is, a one-way service.

Many Twitter-related games exist, but most of the games simply post to Twitter and do not integrate other types of Twitter functionality.

3 THE SYSTEM

3.1 Design policy

The design policy of the service is shown below:

(1) Mutual interaction between Twitter and the service

Few games work with Twitter in a cooperative sense. Therefore, we provide a game that interacts with Twitter mutually.

(2) A user-friendly interface and service

The game must be usable by anyone, regardless of age or gender. Therefore, the user interface must be easy to use.

We developed a service composed of a quiz and the sugoroku game, which are well-known in Japan.

3.2 Development environment

We use the twitteroauth library from the Twitter API [5] and Flash to develop the sugoroku system.

Table 1 shows a list of software that comprises this service.

Table 1: Software constitution.

Component	Software
Web server	Apache
RDBMS	MySQL
Scripting language	PHP, Action Script3.0, Flash

3.3 Function

This service consists of Twitter and a game (sugoroku). Players answer problems in Twitter, and the results are translated as the progress in sugoroku. Figure 1 shows the relation between Twitter and the sugoroku game.

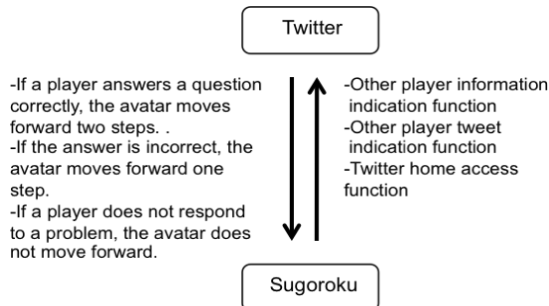


Figure1: Relation between Twitter and the sugoroku game.

(1) Problem contribution function

This function posts the description of the problem and four choices for the problem as one tweet. The problem is randomly selected from a database. Figure 2 shows an example of a problem that was posted on Twitter.



Figure2: An example of a problem that was posted on Twitter.

(2) Right or wrong judgment function

This function judges whether the answer provided by a player is right or wrong. If the response tweet includes the choice number of the correct answer or the words of the correct answer, this function concludes that the player answered it correctly. If the response tweet does not include the correct answer, the function concludes that the answer is incorrect.

(3) Correct answer contribution function

This function posts the correct answer to a problem as a tweet. Figure 3 shows an example of a posted correct answer. The characters of “and 4164” of Fig.3 is an identifier. The number of steps that a player advances in the game is included in the tweet. Figure 4 shows an example of a tweet whether each player posted the right or wrong answer. The number of advancing steps is also included.



Figure 3: An example of a posted correct answer.



Figure 4: An example of a reply indicating whether each player provided the right or wrong answer.

(4) Sugoroku state update function

This function updates the state of the game, including the new position of the players and the number of turns taken by the players. If a player answers a question correctly, the avatar moves forward two steps. If the answer is incorrect, the avatar moves forward one step. If a player does not respond to a problem, the avatar does not move forward.

(5) Goal arrival contribution function

This function posts the name of the player who arrived at the goal and the number of turns taken to arrive at the goal. The post is a normal tweet, not a reply. Figure 5 shows an example of a tweet indicating the arrival of a player at the goal. The characters of “and939” of Fig,5 is an identifier.

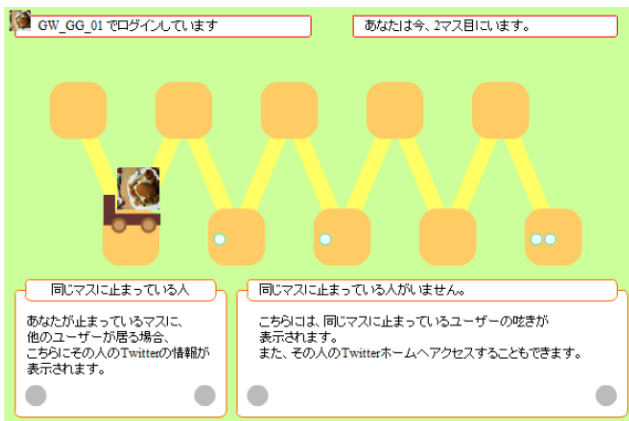


Figure 5: An example of a screen of a tweet generated by the goal-arrival contribution function.

(6) Sugoroku state indication function

This function displays the state of the sugoroku game. Figure 6 shows an example of a screen of a sugoroku game. The avatar of the player is displayed using the player's profile image in Twitter on the load-carrying platform. The avatars of other players are displayed as small circles in the ten steps. Each player starts at the top-left step.

If two or more avatars are in the same step, those players can communicate with each other, and their details are shown, so that they could access each other's Twitter pages.



[双六からログアウトする](#)

Figure 6: An example of a screen of the sugoroku game.

(7) Other player information indication function

This function displays the Twitter information about other players in the same step.

(8) Other player tweet indication function

This function displays tweets from other players in the same step. Figure 7 shows an example of a tweet from another player.

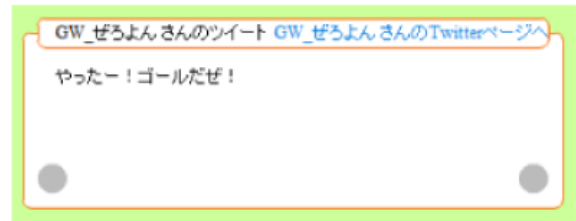


Figure 7: An example of a tweet from another player.

(9) Twitter home access function

This function shows the Twitter home of a player in the same step. Figure 8 shows three players in the same step in the lower-right side. If the player clicks the area indicated by the white arrow, then the Twitter home of that player is displayed (bottom part of Fig.8).



Figure 8: An example of a screen that enables access to another player's Twitter page.

3.4 Operation

To participate in this service, the GameMaster account must be followed in Twitter. The GameMaster account in Twitter posts the questions that players must answer. This system considers any player responding to the GameMaster account to be a service participant. By responding to the GameMaster account, the posted problem is displayed on the player's timeline. Figure 9 shows the flow of the service.

The GameMaster account posts questions regularly. However, the participants of the game must answer the problem within a time limit. The player can answer many times as long as the replies are within the time limit, but

only the last answer of the player is considered the final answer.

After the time limit passes, the GameMaster account posts the correct answer. The player advances on the sugoroku game only for the steps mentioned in the tweet.

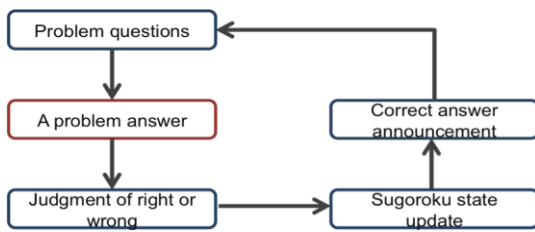


Figure 9: The progress of the service.

The questions are created and posted regularly. One of the questions is shown below:

“What is the public transportation in Seattle?”

The answer is chosen from the options below:

1: A cable car, 2: An underground bus, 3: A subway, 4: A ropeway

2: An underground bus is the correct answer.

Examples of other question are shown in Appendix A.

4 EXPERIMENTS AND RESULTS

4.1 Experiment

Eleven students from Wakayama University were the subjects for our experiment. Most (10/11) have used Twitter. The subjects were asked to use the accounts that we prepared. The subjects were divided into two groups with differences in intervals between questions.

The subjects were free to post tweets during the experiment and to use the service. The subjects do not know the names of the other players. The experiment was conducted in one full day. Figure 10 shows an example of a screen of tweets. A player answers a question. Every player can view the answer. The answers are checked every five minutes. Table 2 shows details of the experiments.



Figure 10: An example of a screen of tweets.

Table 2: The details of the experiments.

	Number of the subjects	Number of the questions	Questions interval (min.)	Answer time (min.)
Exp.1	5	19	60	30
Exp.2	6	32	45	30

4.2 Experimental results

The result of Experiment 1 is shown in Table 3. Only Subject D reached the goal. The number of questions is 19.

Table 3: Result of Experiment 1.

Subject	Number of answers	Position at the end of the game
A	1	2
B	3	6
C	5	9
D	8	10
E	3	4

The result of Experiment 2 is shown in Table 4. Subjects A, B, C, D, and E reached the goal. The number of questions is 32.

Table 4: Result of Experiment 2.

Subject	Number of answers	Position at the end of the game
A	15	10
B	5	10
C	7	10
D	6	10
E	5	10
F	5	8

Table 5 shows the results of a questionnaire about the problems and answers in Twitter. These questions were rated using a five-step scale.

Table 5: Results of questionnaire about the problems and answers in Twitter.

Question	Exp.1	Exp.2
Q1: Was the interval between questions appropriate?	4.0	3.8
Q2: Was the number of the questions within one day appropriate?	2.8	4.0
Q3: Was it easy to answer the question?	4.8	4.3
Q4: Was Twitter pleasant to use for the questions and answers?	4.6	5.0

1: I strongly disagree, 2: I disagree, 3: I neither agree nor disagree, 4: I agree, 5: I strongly agree

Table 6 shows the results of a questionnaire about Twitter and the sugoroku game. These questions were rated using a five-step scale. The evaluation value is the mean for all 11 subjects.

Table 6: Results of questionnaire about Twitter and the sugoroku game.

Question	Evaluation
Q5: Did you become interested in the other players?	3.8
Q6: Did you feel inclined to answer a problem in Twitter by the sugoroku game?	3.8
Q7: Did you feel that the sugoroku and Twitter worked in tandem?	4.2
Q8: Was this service pleasant to use?	4.6

1: I strongly disagree, 2: I disagree, 3: I neither agree nor disagree, 4: I agree, 5: I strongly agree

4.3 Discussion

First, we discuss the problems and answers in Twitter (Table 5).

Q1: "Was the interval between questions appropriate?" The ratings are 4.0 (Exp.1) and 3.8 (Exp.2); therefore, the interval between questions is better around 60 minutes than 45 minutes. But, in the case of Exp.2, everybody aimed at the goal. Experiment 2 (every 45 minutes) may be better to concentrate on a game.

Q2: "Was the number of the questions within one day appropriate?" The ratings are 2.8 (Exp.1) and 4.0 (Exp.2); therefore, we think that more questions are better.

Q3: "Was it easy to answer the question?" The ratings are 4.8 (Exp.1) and 4.3 (Exp.2). Both results are high. We think that the design of the system is good; for example, the reply system using Twitter and the issue of choice.

Q4: "Was Twitter pleasant to use for the questions and answers?" The ratings are 4.6 (Exp.1) and 5.0 (Exp.2). Both results are high. Answering in Twitter is regarded as a pleasant game experience. Because, the answer of other player can be viewed, players can suppose the answers (Fig.10).

The subjects needed knowledge on a wide variety of subjects. A player enjoys the game more if he/she has such knowledge.

Next, we consider the interaction between Twitter and the sugoroku game.

Q5: "Did you become interested in the other players?" The rating is 3.8, which is comparatively high. We think that we can promote interest in other players on the same step by displaying their information on a sugoroku screen.

Q6: “Did you feel inclined to answer a problem in Twitter by the sugoroku game?” The rating is 3.8, which is comparatively high. We could conclude that answering questions in Twitter was encouraged by the sugoroku game.

Q7: “Did you feel that the sugoroku and Twitter worked in tandem?” The rating is 4.2, which is high. We think that the cooperation between Twitter and the sugoroku game are good, based on the result of Q6 above.

The evaluation about Q8: “Was this service pleasant to use?” The rating is 4.6, which is high.

We believe the rating was generally high, because of the players’ knowledge on a wide variety of subjects.

5 CONCLUSION

We proposed a service that uses Twitter in a sugoroku game. We tested this service twice, and the results of experiment indicated that the game was interesting (4.6/5.0) by Q8, and the interaction between Twitter and the sugoroku game was evaluated relatively high (3.8/5.0) by Q5 and Q6. However, we should compare other related games.

In the future, we would like to modify the questions based on the progress of the sugoroku game and to enable participation of more people.

REFERENCES

- [1] <http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h23/html/nc213120.html>
- [2] Twitter: <http://twitter.com/>
- [3] Dotwar, <http://dw.sipo.jp/>
- [4] Furai no shiren 4+, <http://www.spike-chunsoft.co.jp/games/shiren4/>
- [5] abraham/twitteroauth • GitHub
<https://github.com/abraham/twitteroauth>

APPENDIX A

The following are examples of the questions posed to the players:

Question: What is the capital of Australia?
(Answer: Canberra)

Question: What prefecture includes the highest mountain in Kyushu?
(Answer: Kagoshima)

Question: What is the name of the food created with the

water used to boil the jaja noodles of Morioka?
(Answer: Chitantan)

Question: What is the sexagenary cycle of 2009?
(Answer: The Ox)

Question: What is the deep-fried chicken called in Hokkaido?
(Answer: Zangi)

Question: Which company runs the route bus with the longest route in Japan?
(Answer: Nara Kotsu Bus Lines Co., Ltd.)

Question: Which company supplies electricity in Yakushima?
(Answer: YAKUSHIMA DENKO CO., LTD.)

Extension Mechanism for Integrating New Technology Elements into Viewpoint based Enterprise Architecture Framework

Akira Tanaka^{*}, and Osamu Takahashi^{**}

^{*}view5 LLC, Yokohama, Japan

^{**}Department of Media Architecture, Future University Hakodate, Japan
atanaka@view5.co.jp, osamu@fun.ac.jp

Abstract – Enterprise Architecture needs to cover various aspects of the target enterprise, starting from business strategy to communication protocols. However, business environment changes daily, and new technologies are constantly introduced. There is a need to define a basic mechanism for integrating new technologies into Enterprise Architecture. This paper presents a candidate mechanism for integrating recently accepted technology elements, such as mobile/cloud computing, and social network. The proposed mechanism is for RM-ODP viewpoint language based Enterprise Architecture, which includes modification approach to meta-models and UML Profiles. Findings and discussion covers integration approaches, relationship with model driven software development, and openness or interoperability of Enterprise Architecture.

Keywords: Enterprise Architecture, RM-ODP, Mobile Device, Cloud Computing, Social Network, UML

1 INTRODUCTION

1.1 Enterprise Architecture

Enterprise Architecture is a widely recognized term for designing enterprise's software and system architecture. Examples include Zachman Framework [1][2], TOGAF [3], and Federal Enterprise Architecture (FEA) [4]. In standards domain, RM-ODP [5][6][7] is an ISO/IEC/ITU-T standard for specifying Open Distributed Processing, which can also be considered as an Enterprise Architecture Framework. It provides foundational concepts and standard viewpoints including languages and structuring rules for specifying enterprise systems. The viewpoints of RM-ODP include Enterprise, Information, Computational, Engineering, and Technology. Other frameworks use different classification scheme: Perspectives in Zachman Framework, Architecture Domains in TOGAF, and sub-architecture domains in Federal Enterprise Architecture. Because of its neutrality, openness, availability of tools, we chose RM-ODP to represent Enterprise Architecture for use in this paper.

One issue with Enterprise Architecture is that they are designed to be stable, which is good but at the same time that may imply it is hard to modify the architecture or difficult to introduce new architectural elements. Therefore, there is a need to study how Enterprise Architecture can be designed to become flexible enough to incorporate changes in consistent manner.

1.2 Modeling Software Architecture

Widely used modeling language for specifying software systems is UML or Unified Modeling Language [12]. It provides a means for defining structural and behavioral aspects of the target system, using its modeling elements such as Class, Component, and Activity. Although it is a general purpose modeling language, with its profiling mechanism users can define customized modeling elements for specific domains such as embedded systems, real-time systems, and Enterprise Architectures. Example profiles covering Enterprise Architecture domain include UML Profile for DoDAF and MODAF [23] and Use of UML for ODP system specifications [8][9]. It usually starts with defining meta-model for clarifying relevant domain concepts, followed by defining its UML profile. This approach is also used for defining meta-model elements of other types of technologies described in the next clause, so that we can work on integration of those meta-models and UML profiles.

2 RECENT TECHNOLOGIES' IMPACT ON ENTERPRISE ARCHITECTURE

Growing popularity of mobile devices such as smart phones and tablets, cloud computing, and social networks are examples of the changes we have witnessed recently. Although those are just examples, those technologies were not really considered in Enterprise Architecture ten years ago, but we need to integrate them today. We will examine kinds of impact those technologies bring to Enterprise Architecture.

2.1 Mobile Devices

Mobile devices at the time were mainly portable laptop PCs and communication means were limited.

Reason for success may include Moore's law, people's welcome of mobile devices (actually computer systems), and acceptance of living in cyber world (emails, social networks etc.).

Impact on Enterprise Architecture includes addition and processing of new data elements like object's location, position, direction and time-stamp. Additional work of adjusting UI for mobile devices and increased security concerns e.g. by stolen devices or less secure network is also part of the impact. Mobility may be applied to people and things, such as software elements and hardware elements,

specified in Enterprise Architecture. The impact would be to all the viewpoints.

2.2 Cloud Computing

Application Service Provider or ASP might be conceptually close to today’s SaaS (Software as a Service). PaaS (Platform as a Service) equivalent did not exist, and rental server services or hosting services of the time could be considered as services similar to IaaS (Infrastructure as a Service) with limited capability.

Reason for success may be that cloud computing usually provides better ROI of IT resources than internal investment in hardware, software, system administration, training, and system development.

Impact on Enterprise Architecture includes increased use of external application services (SaaS), and external platforms to run the applications (IaaS and PaaS). There are various types of clouds: public cloud, private or internal cloud, hybrid cloud, and personal cloud. NIST’s definition of cloud computing [10] is widely accepted. However, the use of external resources requires strong governance over security and regulatory issues. The impact would be to enterprise viewpoint for external applications and to computational/engineering/technology viewpoints for integration with this technology.

2.3 Social Network

Except for forums provided by e.g. online service providers, there was nothing comparable with today’s social networks in popularity and scale.

Reason for success may be that enterprises realized the importance of people aspect, including e.g. managers, developers and operators in enterprise systems, by looking at its success in consumer market. It is, therefore, logical to consider including this capability into enterprise computing systems.

Impact on Enterprise Architecture includes addition and processing of new data elements (social profile and people oriented network information), which leads to a new class of applications that contributes to better performance by enabling posting, reading, reacting to messages to construct his/her social networks and analyzing social network [11], in the context of enterprise systems. The impact would be to all the viewpoints.

Those three elements, mobility, cloud computing, and social network, can have the following relationship (Fig. 1).

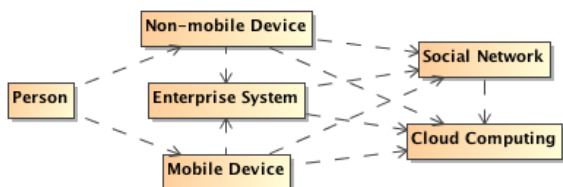


Figure 1: Relationship/Dependencies between mobile devices, cloud computing, and social networks

Mobility enhances access capability to a part of enterprise system running on cloud platform, cloud computing’s scalability and reliability supports dynamic traffic changes

of social networking, and mobile access to social network will accelerate the use of social network.

3 ANALYSIS AND PROPOSED MODIFICATIONS ON RM-ODP

We analyze these three technology areas and consider meta-models or UML Class diagrams representing MOF [13] model, consisting of their core concepts and its relationships among them, and with RM-ODP concepts, in order to define integrated meta-models for UML Profile development. Note that some concepts from three areas may already be defined in RM-ODP, and we will analyze the difference, give priority to existing concepts, and add necessary semantic differences to the model element when needed. Other elements are assumed to be independent and to be integrated into RM-ODP meta-model as domain-specific extensions. This can be explained with the use of UML’s package merge and package import, like the following diagram (Fig. 2).

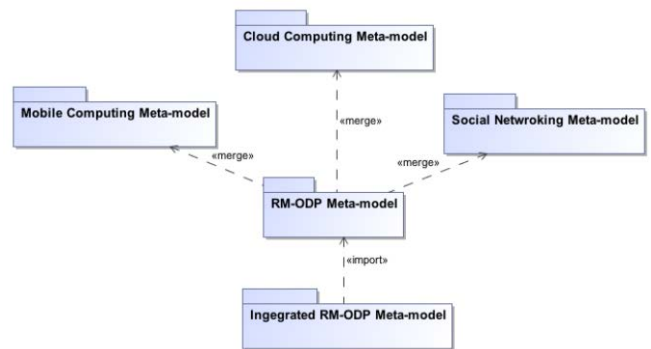


Figure 2: Meta-model integration with package merge

3.1 Mobile Devices

In case of mobile devices, mobility elements were not present in the most Enterprise Architectures. Some did have location concept but it was not meant to be dynamic, e.g. specified department’s location such as “Stockholm, Sweden” does not change every minute. In order to add mobility to an object, it needs to have a mobility attribute or an optional link to mobility. Mobility attribute or mobility consists of time-stamp and location. If necessary, velocity of moving object can be computed using this record ($v = \Delta \text{location} / \Delta t$). There is a prior work on defining UML Profile for Mobile Systems, which shows place or location as important concept to be introduced [14].

Modifications to RM-ODP are the followings.

Enterprise/Computational/Engineering/Technology Languages: Optional link to Mobility added to Viewpoint Objects

Information Language: No change [since information models do not change depending on time and location]

As proposed changes to RM-ODP concepts, Mobility can be defined as a combination of LocationInTime and LocationInSpace (Fig. 3, fragment of the modified meta-model), both are defined terms in RM-ODP. Geographic Information standard could be used for LocationInSpaceType.

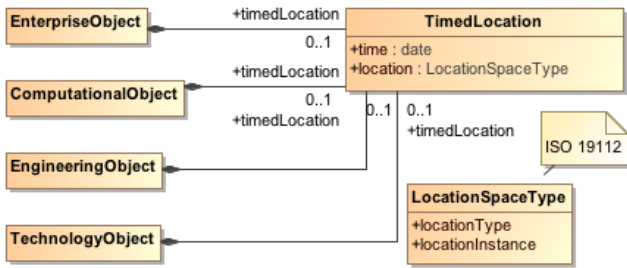


Figure 3: TimedLocation

3.2 Cloud Computing

In case of cloud computing, especially in business domain, the major concerns are actually on business execution, and underlying technologies such as computing platform are considered as engineering issues. However, use of cloud computing based external services can become an impact. When they are incorporated into business processes, they will work as action/activity or step implementations within certain business processes. Those external cloud applications usually provide services using web interfaces. Therefore, to incorporate external services, support of web interfaces or Service-Oriented Architecture (SOA) [15] will be needed in the architecture. In addition, integration with existing systems (legacy systems) could also be done in the same fashion. Interaction with external services, though, will require its policy to cover agreement for using external services and/or regulatory restrictions. NIST has published Cloud Taxonomy as part of its Cloud Computing Reference Architecture, where major cloud computing concepts are introduced [16].

Modifications to RM-ODP are the followings.

Enterprise/Computational/Engineering/Technology Languages: Optional link to CloudService was added to Objects, and a new datatype “CloudServiceType” was introduced.

Information Language: No change [since information models do not change depending on where and how they are managed]

As proposed changes to RM-ODP concepts (Fig. 4, fragment of the modified meta-model), CloudNature can be defined as a combination of CloudSupport (Boolean) with CloudType ((Public, Private, Hybrid) & (SaaS, PaaS, IaaS)). Also in Enterprise Viewpoint, a Step may be labeled with CloudNature, showing that an Object with Cloud Support is supporting the Step.

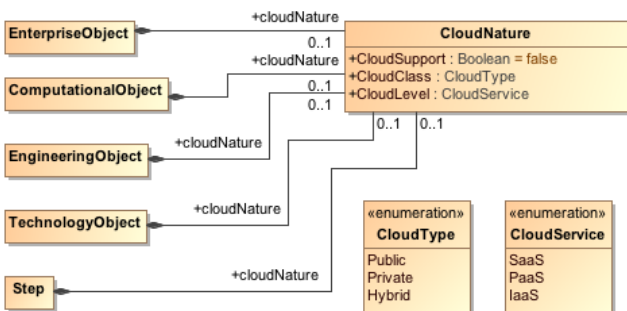


Figure 4: CloudNature

3.3 Social Network

When a person starts working for an enterprise, he/she will be given a title or role in an organization, which is connected to what the person is obligated to do, allowed to do, and prohibited to do, or job description. This model needs to be modified to incorporate sociality, which is described in the person’s social profile (Fig. 5, fragment of the modified meta-model). A new data types such as person’s interests, experience, and participating social communities with roles within need to be there. Person is usually modeled as Party, which can have a relationship with other Parties. RM-ODP’s Community concept is a good fit to represent social community. The resulting architecture will include parties, services, processes, etc. within the enterprise just like normal business processes to make best use of people’s capability. There is a work related to our meta-model, which lists similar concepts that we have here [17].

Modifications to RM-ODP are the followings.

Enterprise Language: Optional link to SocialProfile with SocialRelationship were added to Party, and definition of SocialProfile was added.

Information/Computational/Engineering/Technology Language: No change

We use Party as defined in RM-ODP and introduced SocialRelationship, Social Profile, and Social Community, which is a subclass of Community. We can also use suitable Viewpoint Language elements.

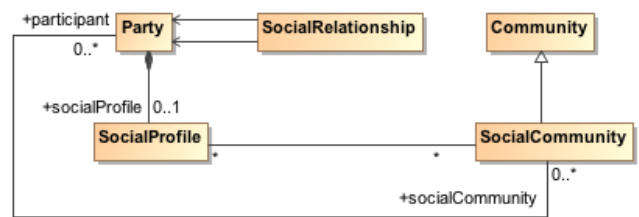


Figure 5: SocialProfile

4 UML PROFILES

The changes to conceptual model need to be reflected in the UML Profile. There is an ISO/IEC/ITU-T standard called “Use of UML for ODP system specifications” or UML4ODP for short, with which ODP models can be created with UML tools.

There are two ways for modifying UML Profile. First is to modify the existing stereotypes, and the second is to define and introduce new stereotypes such as MobileObject and CloudObject by inheriting from existing stereotypes. The latter will lead to a creation of many subclass stereotypes such as NV_MobileBinder and NV_CloudBinder. In order to avoid the too-many stereotypes and the need to introduce new diagrams, we chose modifying existing stereotypes approach.

This integration can also be explained with the use of UML’s package merge and package import. The mapping of modified parts of the meta-models to UML Profiles is discussed below.

4.1 Mobile Device and Cloud Extension as UML Profile

Each viewpoint object’s stereotype (EV_Object, CV_Object, NV_Object, and TV_Object) is enhanced to include attribute definitions covering mobility and cloud-ness (Fig. 6). An attribute “mobility” is a Boolean with default value false, meaning if it is true the object is mobile object. Only in that case, time and location attributes are set. In the same manner, an attribute “cloud-ness” is a Boolean with default value false, meaning if it is true the object is cloud-supported object. In this case, cloud type and cloud service type attributes are set.

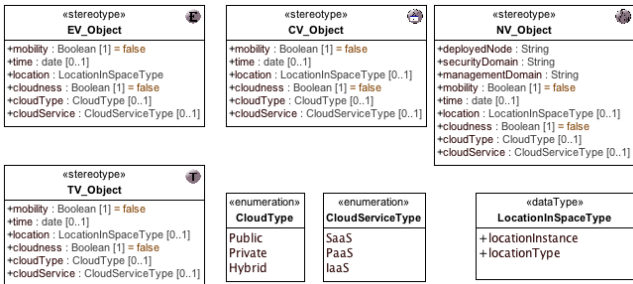


Figure 6: Stereotypes for Mobility and Cloud

4.2 Social Network Extension as UML Profile

EV_Party was enhanced to have sociality, social information, and social communities attribute definitions. EV_Community was enhanced to have sociality and participants attribute definitions. A new stereotype SocialRelationship, which extends UML Association, was also introduced (Fig. 7). The details of SocialInformationType are not defined here.

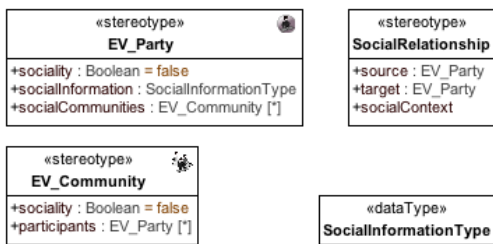


Figure 7: Stereotypes for Social Network

5 IMPACT ON ENTERPRISE ARCHITECTURE ELEMENTS

We have applied those stereotypes together to see what kinds of impact Enterprise Architectures receive.

5.1 Mobile Device

We observe the following impact.

1) Enterprise Viewpoint: The most part of this viewpoint model does not get impact of mobile object, since this viewpoint model mainly talks about why and what. However, a Role, an abstraction of behavior performed by Object, may get influence by mobile object. Especially, policy value will need update to include the cases where

some Roles are performed by mobile object. For instance, if a mobile object performs a part of the process or interaction, e.g. new security policy X may apply.

- 2) Information Viewpoint: No impact
- 3) Computational Viewpoint: Although this viewpoint does not care about distribution, mobility is functionality, and a mobile object can still be introduced. That will give surrounding objects some impact. For instance, an object providing geographical map based on a mobile object’s TimedLocation information may be introduced to support mobile objects.
- 4) Engineering Viewpoint: A case of a mobile object moving from Node A to Node B becomes a possibility. Also, a mobile object may need multiple channels for communication, since available channel may be different from place to place and from time to time.
- 5) Technology Viewpoint: Technology objects representing software, hardware, and network will be categorized into mobile and non-mobile object.

5.2 Cloud Computing

We observe the following impact.

- 1) Enterprise Viewpoint: The same observations as above (5.1 1)) apply. Policy value will need update to include the cases where some Roles are performed by cloud object. For instance, if a part of the process (or a step) or interaction is performed by a cloud object, or an artifact used is fulfilled by a cloud object, e.g. new security policy Y may apply.
- 2) Information Viewpoint: No impact
- 3) Computational Viewpoint: Computational Viewpoint model specifies distributed transparency attributes as a whole (see UML4ODP). Depending on the cloud provider or service, a part of distribution transparency may be provided by the cloud, which means with cloud object computational model may become composite of ODP specified distribution transparency part and cloud provided distribution transparency part.
- 4) Engineering Viewpoint: In case of SaaS, Engineering Viewpoint model including cloud objects and channels to communicate with them is all we need to define. Other elements such as Node for SaaS may be created as virtual element. In case of PaaS, a cloud object providing specific application functionality and the platform are the main elements to be modeled. Other elements such as Node for PaaS may be created as virtual element as well. In case of IaaS, it is possible to model most of the engineering viewpoint except for Nucleus etc.
- 5) Technology Viewpoint: Technology objects representing software, hardware, and network will be categorized into cloud and non-cloud object.

5.3 Social Network

We observe the following impact.

1) Enterprise Viewpoint: SocialParty, SocialRelationship and SocialCommunity are additions to the viewpoint model, and those need to be defined. The behaviors defined in the model will need updates to reflect the new elements. A process to construct social profile,

setup/execute social activity, and to achieve some social objective with the help of social relationships may be added.

2) Other Viewpoints: No impact except for normal viewpoint modeling of supporting viewpoint objects for communicating with social networks.

6 APPLYING PROFILES TO MAJOR ELEMENTS OF ENTERPRISE ARCHITECTURE

We have applied all the UML Profile elements described before into existing UML4ODP Profile definition. With this revised UML4ODP, we can create new kinds of models or diagrams as a step towards Flexible Enterprise Architecture.

6.1 Enterprise Viewpoint Model

There are various types of model or diagram in Enterprise Viewpoint when UML4ODP is used. The following covers only major diagrams.

Objective diagram: A diagram showing Objective decomposition

CommunityContract diagram: A package diagram showing Community and Objective, a package of EnterpriseObjectTypes, a package of Roles, a package of Policies, and a set of Processes.

EnterpriseSpec diagram: A package diagram showing included CommunityContract packages and associated FieldOfApplication.

EnterpriseObjectTypes diagram: A package showing included EnterpriseObjectTypes including ODPSystem, and the relationships among them

RolesInCommunity diagram: A Community and a list of Roles involved

RolesObjects diagram: A diagram showing a set of Roles, a set of Objects, and FulfillsRole relationships between them

Interaction diagram: One or more Interactions with associated Roles, referenced Artifacts, and Enterprise Objects fulfilling the artifact roles

Process diagram: An activity diagram showing Roles and their behaviors (Steps, Artefact, etc.)

Policy diagram: A set of Policy Envelope, Policy Value, relevant controlling Process, and affected behaviors such as Interactions

A package of Enterprise Object Types can include Enterprise Object Types with new attributes (Fig. 8).

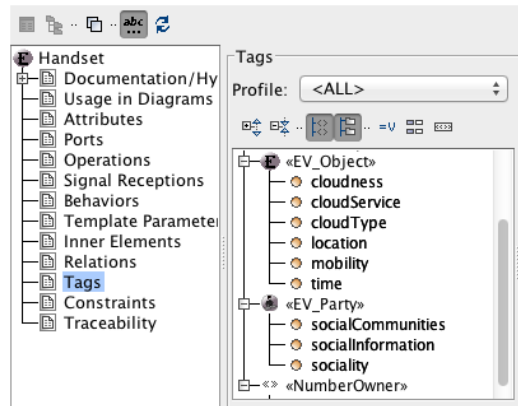


Figure 8: Enterprise Objects with Attributes and Roles

Enterprise Objects with new attributes (see Fig. 8 for Handset) can also appear in Interaction Model (Fig. 9).

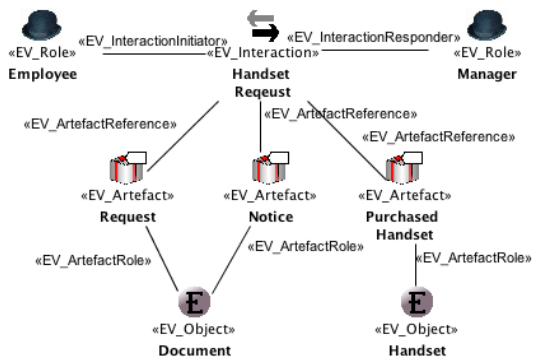


Figure 9: Sample Interaction Diagram

Defined Enterprise Objects can also appear in process diagram (Fig. 10).

Although those are RM-ODP specific diagrams, similar models can be found in other Enterprise Architecture Frameworks.

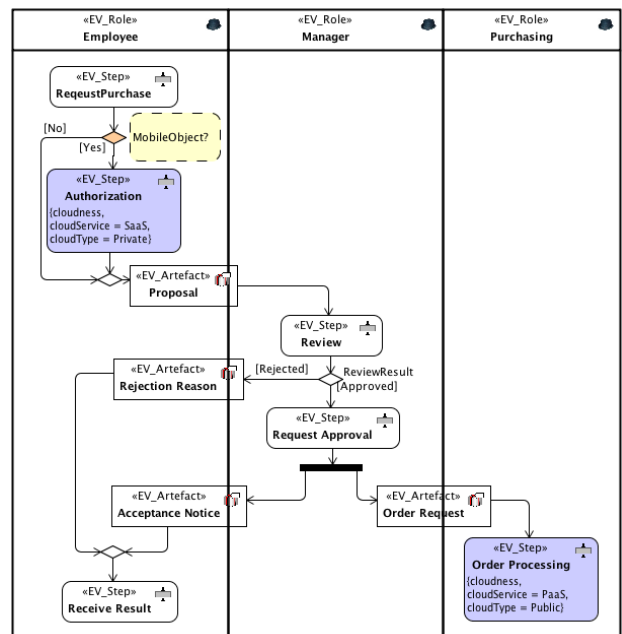
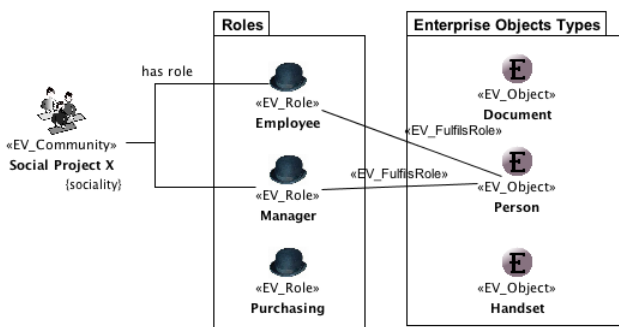


Figure 10: MobileObject and Step in Sample Business Process



6.2 Information Viewpoint Model

We have defined no additional stereotypes for this viewpoint. We can, however, still define additional data types or Information Objects in Invariant Schema diagram using standard UML4ODP (Fig. 11).

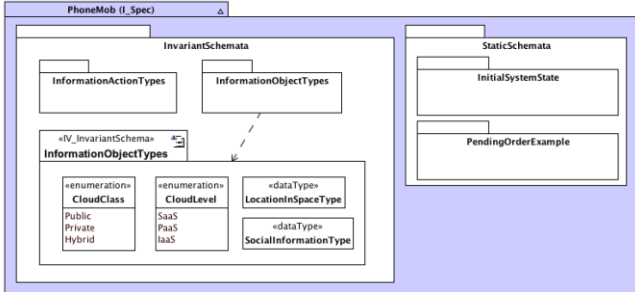


Figure 11: New Invariant Schema Elements

6.3 Computational Viewpoint Model

There are various types of model or diagram in Computational Viewpoint when UML4ODP is used. The followings are two of the examples.

Architecture diagram: A diagram showing logical grouping of architectural packages such as application objects package containing business functions package and ODP function package. Components definition can be a part of this diagram.

Interface/Signature diagram: A diagram showing a set of interface definition and signature definitions, with datatype definitions used.

A Computational Object can have TimedLocation and/or CloudNature as attribute definitions and be used in the architecture diagram below (Fig. 12).

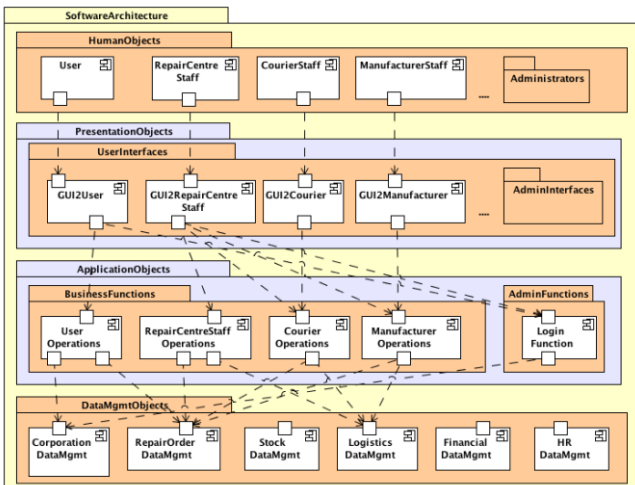


Figure 12: Sample Mobile and Cloud Components

6.4 Engineering Viewpoint Model

Kinds of diagram in Engineering viewpoint are similar to those of Computational viewpoint, and the differences are in distribution-awareness and in internal structure of Node and Channel. When we use e.g. SaaS, it is the objects on remote node to access and use, and in general there is no need or no way to describe internals of the target SaaS system. However for the purpose of this modeling, we used the same stereotypes with attributes (Fig. 13) for describing SaaS object.

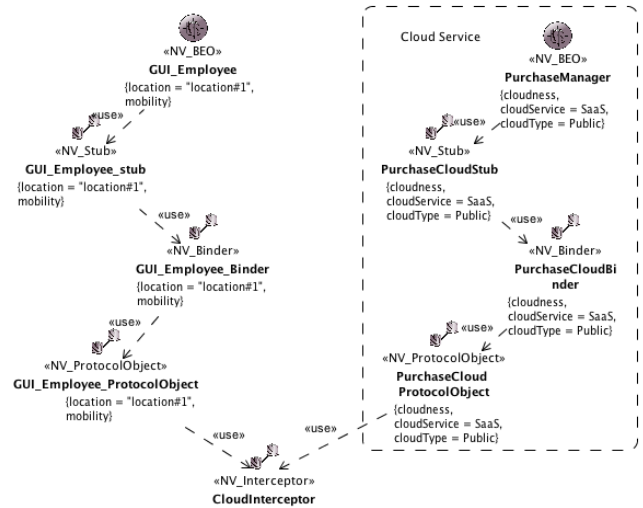


Figure 13: Sample Channel from Mobile Device to Cloud Services

6.5 Technology Viewpoint Model

Technology Object with mobility or cloud-ness can be used to specify elements of hardware, software, or network. However, in case of cloud computing, and if it is a private cloud, there is not much difference with ordinary in house servers case (Fig. 14). However, if it is a public cloud, there is a limit for defining technological architectures, since internal of cloud services is not visible, and cannot be specified, from outside.

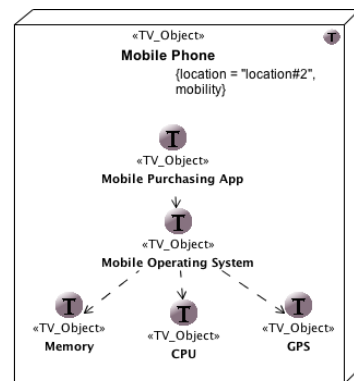


Figure 14: Sample Technology Object for Mobile Device

6.6 Viewpoint Correspondence Model

RM-ODP provides concept of Correspondence, with which we can specify a model element in one viewpoint is related to another model element or model elements in other viewpoint. This is implemented in UML4ODP, and without modifications, we can use this capability in our example as well (Fig. 15).

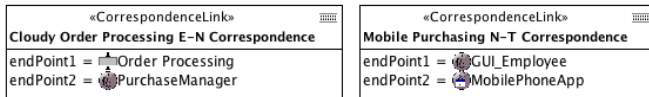


Figure 15: Sample Correspondences

7 FINDINGS IN BRINGING FLEXIBILITY INTO ENTERPRISE ARCHITECTURE

7.1 Summary of Proposed Extension Mechanism

The mechanism we have employed can be summarized as follows. First, we need to have meta-models for Enterprise Architecture and new technologies in the context of Enterprise Architecture. Note it is necessary to understand the domain enough to define meta-models. Second, the meta-model for Enterprise Architecture, as a receiving package, package-merges with one or more of meta-models, as merged packages, for new technologies. When there is a conflict, a process to resolve the conflict described below (7.2, 7.3) should be followed. Third, based on those meta-models, we define UML Profile definitions. UML Profile definitions can also use package merge by treating UML Profile for Enterprise Architecture as a receiving package and UML Profile for new technologies as merged packages.

7.2 Relatively Independent Cases

If target new technologies were relatively independent from existing Enterprise Architecture elements, we should list the overlapping domain concepts, analyze the difference, use the domain concept in Enterprise Architecture as a base concept, and develop additional elements to cover missing semantic elements. There would be at least two enhancement strategies for integration. The first strategy is to add the semantic elements at meta-model level (package merge of meta-models), and enhance existing elements at UML Profile level (package merge of UML Profile). The second strategy is to add the semantic elements at meta-model level (package merge of meta-models), and at UML Profile level (package import of UML Profile). With the first strategy, users will need to understand the updates for each stereotypes, but will be able to use the same set of stereotypes. This can be used in the case that quick implementation is required. With the second strategy, users will need to learn new stereotypes. This can be used to get better modularity of UML Profile, but too many stereotypes may become an issue.

7.3 Conflicting Cases

If the target new technology meta-model has some conflict with existing Enterprise Architecture elements, the conflicting portions need to be resolved. In such cases, meta-model elements from both sides need to be matched in detail to clarify the differences. One such example is SOA concepts against RM-ODP concepts. Key concepts are quite similar but they use different terminologies with the different scope. Similar package merge is used here. Possible resolution for the conflict is to either choose primary one over the other and gradually migrate the second choice to the first one, or to not change anything but treat these as independent and just add correspondence in formal way (e.g. using OCL) or even in informal way. If we need to bring SoaML [18] into UML4ODP, the conflict may be around Interaction (RM-ODP) and Collaboration (SoaML). In this case, choose the one that is more suitable for defining behavior. Another example is BMM (Business Motivation Model) [19] and the conflict is between Objective (RM-ODP) and Vision/Goal/Objective (BMM). In this case it would be less difficult to resolve the case, since BMM provides wider and richer model. It just depends on how deep objective model you need.

8 DISCUSSIONS

8.1 Flexible Extension Mechanism

The question we started with was “is it possible to design flexible or extendable Enterprise Architecture, preparing for the time new technology emerges?” We developed an extension mechanism described in 7.1 Summary of Proposed Extension Mechanism. In the final step of merging UML Profiles, we need to add/modify stereotypes, attribute definitions, and constraints of UML4ODP. When doing this, use of modeling tool is helpful, especially when you already have meta-model data to be revised and profile data to be revised. Modeling tool may also provide validation capability. Even without tooling, it is still possible to take the same steps. However, if you did it by hand, it would be hard to avoid errors and inconsistencies.

8.2 Enterprise Architecture and Model based Software Development

Enterprise Architecture usually means high-level description of the entire enterprise system. That may not change, but if Enterprise Architecture is presented as e.g. a UML model, we can at least make consistent modifications to the model with the help of UML tools. When the idea of Model Driven Architecture (MDA) [20] was introduced, there was no MDA tooling available. Today, there are some commercial and open-source products for model transformations, e.g. eclipse modeling project. Once source model is prepared, it can be used as an input to model transformation tool chains. Enterprise Architecture in UML is a reasonable starting point for this process. Also, if it is UML model, whole or a part of the model may become

candidate for reuse. The real challenge would be to achieve step-by-step model-to-model transformation chains. There are several UML tools that support RM-ODP, Zachman Framework, and TOGAF (e.g. MagicDraw). With those tools and with enhancement support mechanism in place, UML model representing enhanced Enterprise Architecture can be created for consumption by the tool chains. In this case, development of model transformation logic is required but once done it may be possible to reuse, since it is built against standard UML models. Another possibility is use of Domain-Specific Language or DSL [21]. Some people prefer DSL to general purpose UML for its simplicity. If a DSL is designed to describe Enterprise Architecture, then created model in XMI form can also be used as an input for the tool chains. In this case, however, development of model transformation logic is required for each DSL and reuse may become an issue.

8.3 Interoperability among Enterprise Architecture Frameworks

As of today, there is no interoperability among different Enterprise Architecture Frameworks, which produced silos of Enterprise Architectures. The required actions towards interoperability are to make their meta-model open and encourage development of transformations between them, or to make one of the meta-models as standard and each framework provider to develop and provide transformation to/from the standard. The issues in doing this are in resolving the difference in scope or coverage of the concepts in different Enterprise Architecture. The first thing to do is to agree on common core set of concepts for Enterprise Architecture framework. Among the three frameworks mentioned in this paper, RM-ODP based one is the most neutral and open, and this could be used as a base for the discussion on the common core.

9 CONCLUSION

In order to integrate new technologies into Enterprise Architecture, we will need to take the following steps: analyze the domain that new technology is applied and follow the steps described in 7.1 Summary of Proposed Extension Mechanism. The use of “package merge” is for getting flexibility: it allows merging only necessary set of packages into Enterprise Architecture package. The steps handling UML Profile may be replaced with other steps, e.g. generating DSLs using eclipse modeling project, for non-UML model case.

It is likely that new technology provides new capability to things or persons. In this case, new meta-model element related to the capability should be related to Object or Party (Person) so that the capability is explicitly visible to UML Profile designer.

If the Enterprise Architecture model is used in model driven environment as an input file, it should be interpretable by model transformation engines. This means input file should better be in the form of UML or XMI [22].

Regarding openness and interoperability of Enterprise Architectures, we will need a chance to discuss and find

common core concepts. Until this is done, RM-ODP based Enterprise Architecture would be the most open one, since it is an ISO/IEC/ITU-T standard, and standard document, meta-model data, and UML Profile data are available on the Internet.

REFERENCES

- [1] Zachman, J.A., “John Zachman's Concise Definition of The Zachman Framework™,” <http://www.zachman.com/about-the-zachman-framework>, April 2013.
- [2] Zachman, J.A., A Framework for Information Systems Architecture. IBM Systems Journal, Vol 26 (3) 1987
- [3] The Open Group, “TOGAF Version 9.1,” <http://pubs.opengroup.org/architecture/togaf9-doc/arch/>
- [4] FEA Reference Models, Consolidated Reference Model Version 2.3, http://www.whitehouse.gov/sites/default/files/omb/assets/fea_docs/FEA_CRM_v23_Final_Oct_2007_Revised.pdf
- [5] ISO/IEC 10746-2:2009, Information technology -- Open distributed processing -- Reference model: Foundations
- [6] ISO/IEC 10746-3:2009, Information technology -- Open distributed processing -- Reference model: Architecture
- [7] ISO/IEC 15414:2006, Information technology -- Open distributed processing -- Reference model -- Enterprise language
- [8] ISO/IEC 19793:2008, Information technology -- Open Distributed Processing -- Use of UML for ODP system specifications
- [9] Hashimoto, D., Tanaka, A., and Yokoyama, M. “Case study on RM-ODP and Enterprise Architecture,” Eleventh International IEEE EDOC Conference Workshop (EDOCW'07), 2007
- [10] National Institute of Standards and Technology, “The NIST Definition of Cloud Computing,” Special Publication 800-145, 2011
- [11] Dreyfus, D and Iyer, B, “Enterprise Architecture: A Social Network Perspective,” 39th Hawaii International Conference on System Sciences, 2006
- [12] OMG, Unified Modeling Language™ (UML®), <http://www.omg.org/spec/UML/>
- [13] OMG, Meta Object Facility (MOF) 2.0 Query/View/Transformation Specification, <http://www.omg.org/spec/QVT/1.1/PDF/>, January 2011
- [14] Grassi, V., Mirandola, R., Sabetta, A., “A UML Profile to Model Mobile Systems,” <<UML>> 2004 - The Unified Modeling Language. Modelling Languages and Applications, Lecture Notes in Computer Science Volume 3273, 2004, pp 128-142
- [15] OASIS, Reference Model for Service Oriented Architecture 1.0, <http://docs.oasis-open.org/soa-rm/v1.0/soa-rm.pdf>
- [16] NIST, Special Publication 500-292 Cloud Computing Reference Architecture

- [17] H. Van Dyke Parunak, James J. Odell, "Representing Social Structures in UML," Agent-Oriented Software Engineering II, Lecture Notes in Computer Science Volume 2222, 2002, pp 1-16
- [18] OMG, Service oriented architecture Modeling Language (SoaML®)
- [19] OMG, Business Motivation Model (BMM), <http://www.omg.org/spec/BMM/>
- [20] OMG, Model Driven Architecture ® (MDA ®), <http://www.omg.org/mda/>
- [21] Fowler M., Domain-Specific Languages, Addison-Wesley, 2011
- [22] OMG, XML Metadata Interchange (XMI®), <http://www.omg.org/spec/XMI/>
- [23] OMG, Unified Profile for DoDAF and MODAF (UPDM), <http://www.omg.org/spec/UPDM/>

Session3:

Communication

(Chair: Tomoya Kitani)

Disaster-Relief Training System Using Electronic Triage with Voice Input

Misaki Hagino[†], Yoshiaki Ando[†] and Ken-ichi Okada[‡]

[†]Graduate School of Science and Technology, Keio University, Japan

[‡]Faculty of Science and Technology, Keio University, Japan
{misaki, ando, okada}@mos.ics.keio.ac.jp

Abstract - When a large disaster happens, medical workers need to determine the priority of medical treatment and do triage first in order to save more people with the limited health resource available. However they cannot use both hands because of writing triage result on triage tags. This hinders the necessary medical treatment that they must conduct in parallel, and slows down the speed of triage. In addition, since the training is very important for medics to get accustomed to disastrous situation, conducting disaster-relief training is of great importance, but frequent training is impractical as the reproduction of a real disaster site is difficult. In this study, we propose an input interface of the triage where medics get information through a monocular HMD (Head Mount Display) and control the system via voice input, so their hands are free to conduct necessary treatment even while doing triage. Moreover, our system can reproduce the situation through the HMD and allow sharing the training status. The evaluation shows that our system enables to perform triage quickly and accurately with medics' hands free and make disaster-relief trainings more meaningful.

Keywords: First-Triage; Monocular HMD; Voice Input; Disaster-Relief Training; Augmented Reality

1 INTRODUCTION

When a large-scale disaster occurs, a large number of people are injured at the same time, so medical workers first need to do triage in order to decide the priority for treatment depending on their severity of injury and the urgency of treatment. The purpose of triage is to save more lives through efficient use of a limited medical resource. To determine the priority of medical treatment for injured people, it is essential to have deep knowledge and experience related to health care.

Currently, paper triage tags are often used to indicate the condition of injured people and to record information of their injuries. However, in a confused situation, it may be difficult to record information due to loss of paper tags attached to the injured people and it may be impossible to respond and adjust quickly to fast changing conditions. In addition, writing something on paper tags is also one of the factors that disturb diagnostic actions from busy medic's hands.

Therefore, in recent years, electronic triage systems have been researched in order to support lifesaving emergency activities. The goal of electronic triage systems is to manage injured people information in real time by digitalizing paper tags and utilizing a wireless sensor network. We have been developing an electronic triage system where injured people information is sensed by electronic tags and is presented

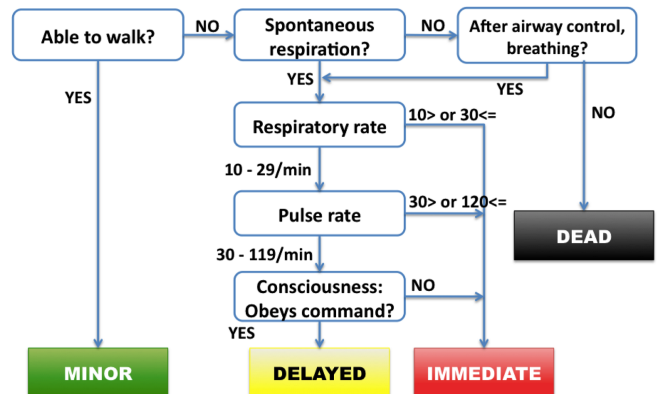


Figure 1: START method

to PDA (Personal Digital Assistance) terminals which health care workers hold.

In this study, we have developed an input interface of the triage where medics can get information such as injured vital signs through a monocular HMD (Head Mount Display) and can control the system via voice input, so that medics' hands are free to conduct necessary immediate medical treatment even while doing triage.

In addition, training is very important for medics to be accustomed with real disaster situations and to know how the electronic triage system works before it is actually used. Our system can reproduce realistic situations through monocular HMDs and allow sharing the training status among participants.

In Section 2, we provide outline and discuss issues of the triage and the training. In Section 3, we show our proposal of solving the issues. Section 4 explains the implementation of our interface. Section 5 discusses the assessment of our system. Finally, section 6 presents our conclusion.

2 CURRENT STATE OF TRIAGE AND TRAINING

2.1 Medical Services Based on Triage

In Japan, medics perform the first-triage based on START (Simple Triage and Rapid Treatment) method shown in Figure 1 at a triage post [1]. The purpose of the first-triage is to classify injured people according to their vital signs such as respiratory rate and pulse rate, so medics only perform 'airway control' and 'airway control' as immediate medical treatment. Injured people are categorized for their treatment priority into the following four: Red (immediate) → Yellow (delayed) →

Figure 2: Paper triage tag

Green (minor) → Black (dead). After each triage, a medic writes necessary information, such as person's name, age, blood type, transportation, organization and the name of the medic in charge, on a paper tag shown in Figure 2, removes unnecessary colors, and attaches it to the injured person [2].

After the first-triage, injured people are taken to each tag color's post and medics perform the second-triage in order to determine priority with transportation to hospitals. In this paper, we focus only on the first-triage.

2.2 Current State of Disaster-Relief Training

After the Hanshin-Awaji Great Earthquake in 1995, residents of the surrounding area have had more opportunities to participate in disaster training with medics. Furthermore, since JR Fukuchiyama Line derailment accident, the concept of triage in lifesaving emergency activities has been widely recognized, and triage training has become essential in order to perform actual triage quickly and accurately in a real situation [3]. Triage training process and its procedure may differ by medical centers, but the basic procedure is as follows. People who play injured role have a piece of paper describing their imitating symptoms and biological information. They need to pretend the written symptoms to medics. Then, medics determine the priority for treatment by observing the behavior of an injured role and information on the paper. After that, medics writes necessary information on a paper tag and makes an order to transport the injured role to the corresponding color triage tent.

For example, Fuji city has conducted a training of doing triage, conducting treatment, recording information and transporting injured people depending on the result of triage [4]. Yamanashi School of Medicine has also conducted an experiment of an electronic triage system, TRACY, where FeliCa IC cards are used [5]. The system aims to share widespread disaster information. In addition, the Emergo Train System, which is a disaster training system with desktop simulation, has also been employed in recent years [6]. Using this training system, medics can study proper arrangement of personnel by moving magnets instead of injured people and medics on a whiteboard which represents a disaster site and hospitals.

2.3 Related Work and Issues

In recent years, systems based on RFID (Radio Frequency Identification) tags and sensors have been researched to be used for emergency lifesaving in disasters [7]. In the Code-Blue Project [8], vital signs obtained from sensors are collected through an ad hoc wireless network. MEDiSN [9] is a sensor network platform for automating physiological monitoring of patients in hospitals and in mass disasters. Furthermore, studies that collect injured vital signs using mobile information terminals, such as PDA (Personal Digital Assistance), have been conducted [10], [11].

However, it is difficult to get vital signs in real time using RFID. Moreover, in order to check information and to enter information, medics have to hold the PDA on their hand. Therefore, they cannot use any hands prohibiting them from performing necessary immediate medical treatment in parallel. In addition, there are other problems in current training systems. In the case of current training systems, biological information of injured people does not change as an actual disaster situation, because medics make diagnosis based on written paper. Moreover, medics perform treatment action and transportation action according to the given manual. An additional problem is that frequent training is impractical, because it requires participation of many medics and people acting injured people, and also requires considerable time and effort to create a scenario of the training and setting up the equipment. The Emergo Train System allows for frequent training because it does not need any actors to play patients, but it does not reproduce a real disaster site or gives a sense of reality. Moreover, the existing training systems cannot utilize triage records or action histories.

3 DISASTER-RELIEF TRAINING SYSTEM FOR ELECTRONIC TRIAGE USING VOICE INPUT

We have been developing an electronic triage system where changes in symptoms and biological information of injured people can be monitored in real time [12]. However, medics need to use their both hands to obtain information such as respiratory rate and pulse rate from the system and use again their both hands to enter triage result on PDA. This hinders the necessary immediate medical treatment that they must conduct in parallel, and slows down the speed of triage. Therefore, we need to develop an input interface of the first-triage system where medics can perform triage with their hands free. In addition, disaster-relief trainings need to be done frequently in order to make effective use of the developed system in real disaster situations, but sharing the status of training is insufficient in the current training system. Therefore, we need to address these issues in our electronic triage training system.

3.1 The Voice Input in the First-Triage

It is desirable to conduct triage in less than one minute per injured person. However, in a chaotic situation, there is a possibility that medics may make a mistake to determine the priority, or may forget the detail of the START method. Fur-

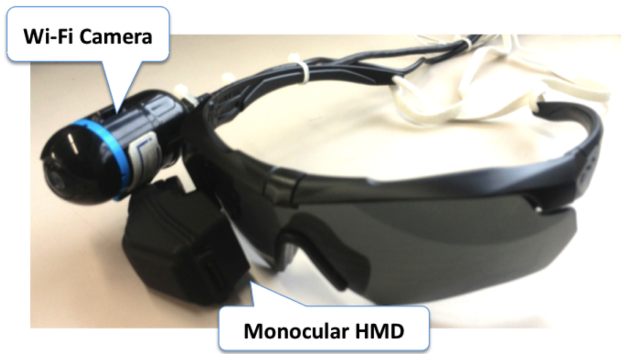


Figure 3: Monocular HMD and Wi-Fi camera

thermore, they cannot use any hands while writing paper tags or using PDA, and cannot fully perform treatment actions allowed for the first-triage, such as ‘astriction’ and ‘airway control’. As a result, the first-triage may take more time.

We have designed voice input function with which medics can perform triage quickly and accurately while keeping their hands free. In addition, we use Vuzix’s TacEyeLT as our monocular HMD, Ai-Ball’s Trek as our Wi-Fi camera to recognize AR (Augmented Reality) markers attached to injured people, and Apple’s iPhone4S as our terminal for screen output to the HMD (Figure 3). Medics operate our system by answering questions displayed on the HMD screen. The questions are also offered as sounds through the earphone so that they can operate without relying wholly on the screen. Answers to the questions are either two choices such as ‘Yes’ or ‘No’ or simple choices which they can answer easily. Moreover, the questions are customized according to each injured person’s vital signs that are automatically collected by the sensors. These functionalities reduce the burden of the thinking of the medics and their mistakes of triage.

3.2 Reproducing Disaster Situation

In triage based on the START method, biological information such as respiratory rate, pulse rate and SpO₂ (oxygen density in blood) become the key element to determine the priority. Therefore, it is necessary to prepare those biological information upon conducting a training. However, it is difficult to ask persons who are actually injured to participate in a training. Although biological information can be gathered from healthy participants, their values are normal and cannot be used for the training to detect any abnormality. In a current training, abnormal values of biological information are written on a piece of paper and triage is performed by referring to those static information. However, actual biological information changes constantly. It is sometimes necessary that some play a role of indicating a sudden change to symptoms. Therefore, an electronic triage training system is desirable which can reproduce situations where biological information is constantly changing.

In addition, it is important in a real diagnosis to see the positional relationship between a injured person and a medic. A medic approaches an injured person, touches the body directly and checks the detail of injury. In order to make a

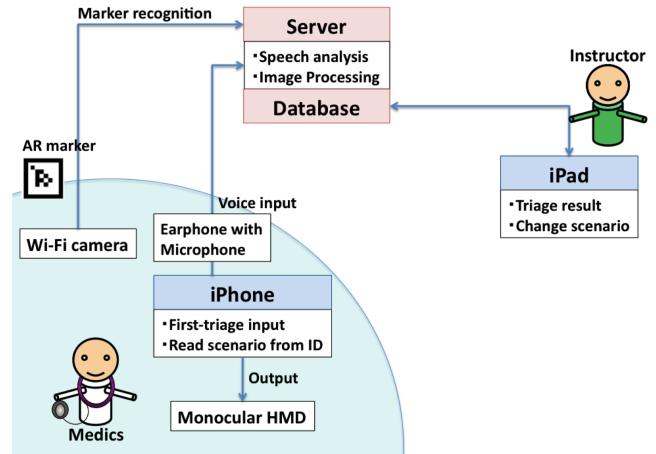


Figure 4: Configuration of the system

diagnosis correctly, a medic needs to see the body closely. Therefore, we need to reproduce the state changes associated with the positional relationship in the training system. In our system, the distance between a medic and an injured person and the direction are used to change how much information is shown.

3.3 Sharing of Training Result

One of the important things in training is to repeat basic trainings [13]. Especially, it is important that medics can find their weak points by themselves after each training is done. However reviewing the result of training is really hard because current training systems use a whiteboard to share training logs such as progress of triage, injured people’s biological information, medics’ actions, and so on. Only those who wrote texts on the whiteboard may understand them. The training is more effective when it is carried out with a group of people. After the training, it is important to review the training logs as well as to understand the situation of other participants. Reviewing the action history after the training enables to analyze mistakes of triage.

4 IMPLEMENTATION

4.1 System Configuration

Figure 4 shows the overall configuration of the system. A medic and an instructor use the system as a pair. The medic who performs triage is equipped with an iPhone and a monocular HMD to which an earphone-mike and a Wi-Fi camera are attached. Sight of user’s one eye is not always be completely blocked, and users can see the HMD screen by looking at the upper part of the glasses. The instructor evaluates medic’s training using an iPad.

First, in the training site the instructor places AR markers which are used instead of injured people. When the medic looks at an AR marker via the Wi-Fi camera, the scenario of the patient corresponding to its ID number can be read. Beforehand, the instructor sets up each injured person’s biological information such as vital signs and the tag color, as that

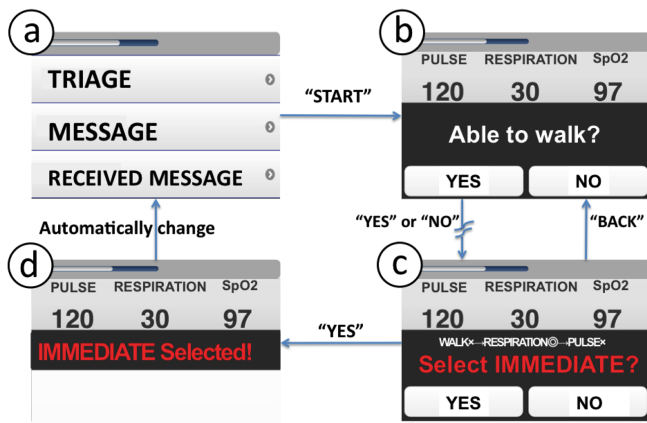


Figure 5: First-triage input flow

person's scenario. During the training, the biological information simulating a real injured person is generated depending on the tag color, and the information management server keeps updating the information at regular intervals. When viewing the AR marker, the medic can carry out triage with voice input by referring injured person's information in the monocular HMD. After the first-triage result is stored, medics can check the tag color by looking at the AR marker again. Via the network, the database stores scenario information created by the instructor, the voice inputs from the medic, the result of triage, the result of marker detection and the result of monocular HMD output. This database enables output of the training result in real time.

4.2 Input Interface of the First-Triage

Figure 5 shows the input interface of the first-triage. The bar on the top of the screen shows the best timing of voice input. Underneath the bar, medics can see respiratory rate, pulse rate and SpO₂. The questions for determining the medical treatment priority based on the START method are displayed in the middle of the screen, and medics can answer the questions using voice input. We have designed each item so that it can be recognizable on a small monocular HMD.

If the camera recognizes an AR marker in screen (a), it will move to the state where triage can be started. If the medic says "start", it changes the screen as is shown in (b), and the medic can answer questions one by one. The system judges automatically the value of vital signs acquired from the injured person, and the next question is selected. Then, the triage result is sent to the database, and the HMD displays the initial screen shown in (a). If the medic makes a mistake, he/she can say "back" and the triage can be redone from the beginning. Thus, the medic can conduct triage without using any hands, and necessary immediate medical treatments can be performed in parallel. We have designed this input interface so that it can be used in a real disaster scene. In the next section, we will explain the case where the input system is used in our disaster-relief training system.

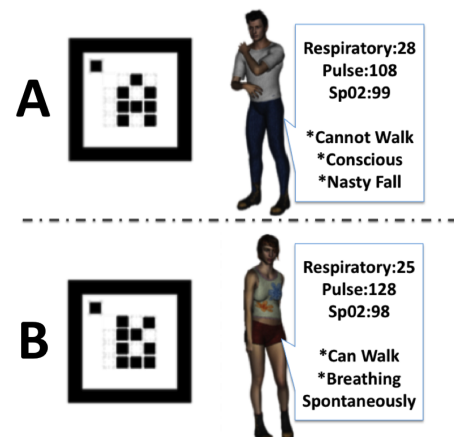


Figure 6: Different scenario depending on ID numbers

4.3 Training System by Reproducing Disaster Situation Through the Monocular HMD

In our training system, injured people are represented by AR markers in a training space. Therefore, the preparation is easy: placing AR markers in the training space. When a marker is recognized by a Wi-Fi camera, the patient's scenario can be read according to its ID number (Figure 6). The reasons why we use AR markers are: anyone can create them easily, and they prevent from misreading ID numbers. We use AR markers looking like alphabets.

The presentation of injured person's information changes according to distance and direction of markers. When performing triage, a medic should face the injured person. Therefore, we reproduced the difference of face-to-face state and side state by the orientation of markers (Figure 7).

Also Figure 7 shows the states depending on the distance from markers. For example, triage can be done only when the distance is less than one meter. According to the distance, the displayed injured person's information changes as follows:

- One marker within one meter: can perform triage and can see all the injured person's information.
- More than one marker or one marker beyond one meter: cannot perform triage and can see only some information of injured persons. The scene is displayed from afar and a diagnosis cannot be made.
- Marker outside of view: cannot triage and cannot see injured person's information.

If the medic says "start" in the bottom screen of Figure 7, interface of the first-triage input is displayed at the right side of the screen (Figure 8).

4.4 Generation of Simulated Biological Information

Our system generates simulated vital signs such as respiratory rate, pulse rate, and SpO₂ dynamically as Table 1 according to each tag color. The values have been determined from

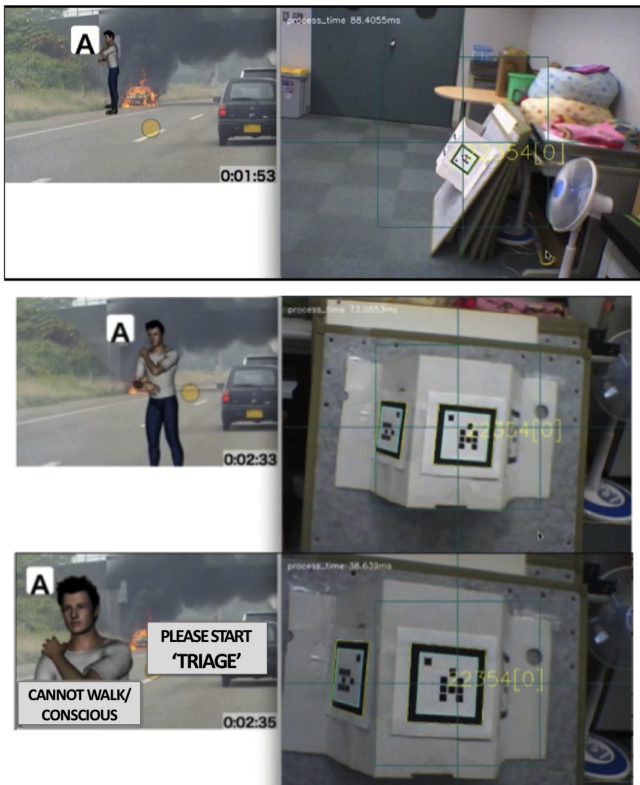


Figure 7: Difference by distance and orientation from markers

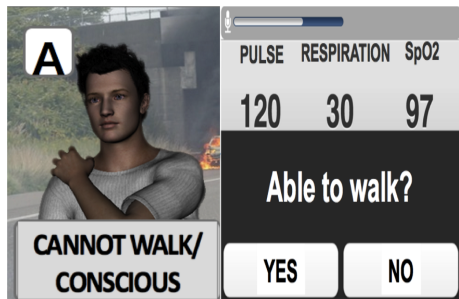


Figure 8: Interface of the first-triage input during training

Table 1: Vital sign parameters

Triage color	Consciousness	Parameter		
		Respiratory	Pulse	SpO ₂
Red	Yes or No (Pattern I)	under 10 or over 30	50-180	90-99%
	Yes or No (Pattern II)	1-50	under 50 or over 120	90-99%
	No (Pattern III)	1-50	20-180	90-99%
Yellow	Yes	10-30	50-120	90-99%
Green	Yes	10-30	50-120	90-99%
Black	No	0	0	0%

the discussion with emergency medical specialists at Juntendo University School of Medicine.

The vital signs are generated at random by our system. We set the upper limit of respiratory rate to 50 breaths per minute and pulse rate to 180 beats per minute. We also set the delta per unit time. Breathing rate's delta is less than 10 breaths per minute, pulse rate's delta is less than 20 beats per minute

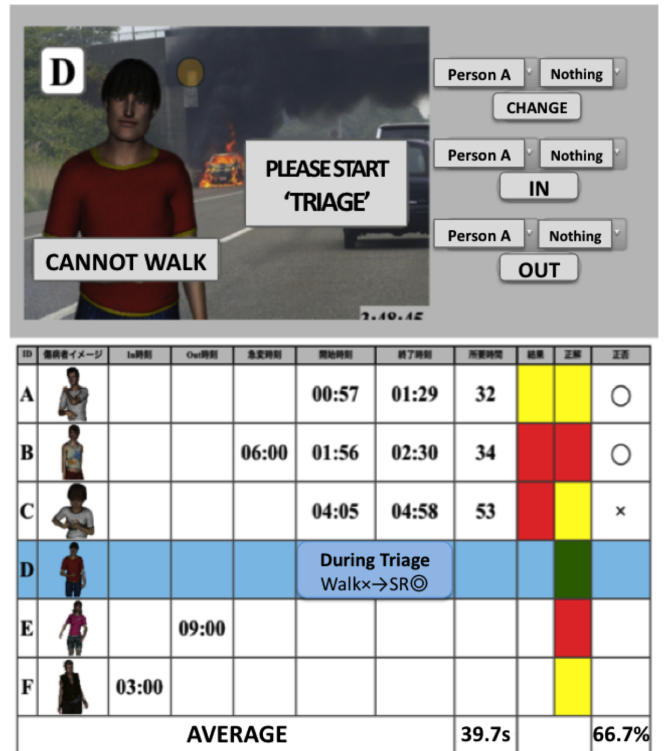


Figure 9: Interface of training result

and SpO₂'s delta is less than 1 percent per minute. This is to eliminate the impractical values and changes in a real situation. We have three patterns which generate vital signs depending on injured people's condition in a case that the triage color is red. Pattern I is when the breathing rate is abnormal and pattern II is when the pulse rate is abnormal. Pattern III is when injured people are unconscious. This provides medics with criteria by means of biological information to decide who need to be given priority for treatment and transportation among multiple numbers of casualties with red tag. In our system, when injured person's condition changes from green or yellow tag to red tag, it is defined as 'sudden change'. If there is sudden change to somebody, medics can recognize it via the monocular HMD.

4.5 Interface of Training Result

After a training, reflecting on one's own actions is very important. Figure 9 shows the display of the training history as seen on the iPad by the instructor. The instructor can see information such as injured person's image according to its ID, when each injured person was carried, when each injured person was transported, when each injured person got sudden change, when triage started, when it finished, how long it took, and how many mistakes were made. The instructor can look back on the flow of triage and the cause of mistakes based on this information. In addition, since the information will be updated in real time during the training, the instructor can continuously monitor how each medic is performing. The top left of Figure 9 shows the monocular HMD display as seen by a medic. The top right shows buttons to change scenario during training. By using these buttons, the instructor

Example 1
<ul style="list-style-type: none"> • Age: 25 • Sex: F • Circumstance of injury: She cannot move by herself. However, she can respond to your call. • Bleeding area: No • Respiration rate: 15/min • Pulse rate: 82/min

Figure 10: Example of an injured person

Table 2: Result of the first-triage input

	Paper Tag	Our System
Required Time (sec)	67.7	32.7
Correctness (%)	86	97

can add other injured people or delete current injured people and make sudden change to arbitrary injured people.

5 EVALUATION

5.1 Evaluation of the First-Triage Input Interface

5.1.1 Procedure

In the first evaluation, we compare our first-triage input system by voice and the current system of using paper based tags. Each examinee performs ten times of the first-triage based on the example of the injured people shown in Figure 10. In the experiment, person type sleeping-bags are used instead of injured people. An examinee performs ‘airway control’ by raising the injured head, and ‘astriction’ by connecting bleeding part with a string, and ‘check of consciousness’ by talking to the injured person. A sheet of paper on which injury is written (i.e. Figure 10) and an AR marker are attached to each injured person (i.e. a sleeping bag). In the case of paper tag system, we asked to write the injured person number, the start time, examinee’s name, the tag color, and the diagnostic flow on the paper tag. In the case of using our system, the system records these items automatically. We selected ten students as examinees. Five of them started off with the paper tag system and then used our system, and the other five did in the reverse order. Because we wanted to compare the paper tag system with our system by the first evaluation, we did not consider the distance and the direction with the AR marker. We wanted evaluate whether examinees were able to perform first-triage using the monocular HMD with voice input.

5.1.2 Result

Table 2 shows the results. Firstly, we focused on the ‘required time’ for input the first-triage. When our proposed system was used, the required time for triage per person was shortened for 35.0 seconds compared with the paper tag. In addition, it is desirable to conduct first-triage in less than one

Table 3: Questionnaire result of the first-triage input

Question	Score
Was the first-triage input easy?	4.5
Were you able to perform medical treatment in parallel?	4.0
Was the screen on the monocular HMD easy to understand?	3.7
Did you feel tired?	3.8

minute per injured person. with our system the examinee was able to perform first-triage in 32.7 seconds. In the case of using paper tags, it was hard to carefully choose selection conditions of the START method and it also took time to record many things on paper tags. In the case of our system, examinees only needed to answer to given questions, and could perform the triage easily. In addition, they were able to perform medical treatment in parallel.

Secondly, if we focus on the ‘percentage of correct answers’, our system is 11 point better than the system using paper tags. The main reason of the improvement is that our system supports examinees to determine the priority by only judging injured vital signs so that they may not need to remember the START method fully.

Furthermore, Table 3 shows the result of the questionnaire about usability. In the evaluation, five is best, one is worst in each item. The questionnaire result shows the first-triage input was easy and hands free was realized. For those reasons, we have confirmed the usefulness of our input interface for the first-triage.

5.2 Evaluation of Simulation Disaster-Relief Training

5.2.1 Procedure

We examined whether examinees were able to easily construct training environment and perform simulation training while wearing the monocular HMD. Examinees were put into pairs consisting of a medic and an instructor, and carried out preparing a training environment, responding to sudden injured person’s condition changes, transporting injured people, and performing triage. The disaster scenario used in the evaluation was a traffic accident involving two automobiles, and seven people were injured. There were two persons with red tag, three persons with yellow tag, two persons with green tag, and none with black tag. These data are taken from a report of the triage training conducted at the Urayasu Hospital of Juntendo University School of Medicine.

Before training, the instructor prepares the training environment by placing AR markers according to Figure 11 for the scenario prepared in advance (event A-1). Once the training started, the medic perform voice input triage while looking at injured people and their information displayed on the screen through AR marker recognition by the Wi-Fi camera attached the monocular HMD (event B-1). During the training, the instructor used the iPad to generate events following the given temporal sequence. If the medic noticed the added events, he/she might respond to the events.

- At 3 min: The instructor adds one injured person with red tag and another with yellow tag (event A-2). The

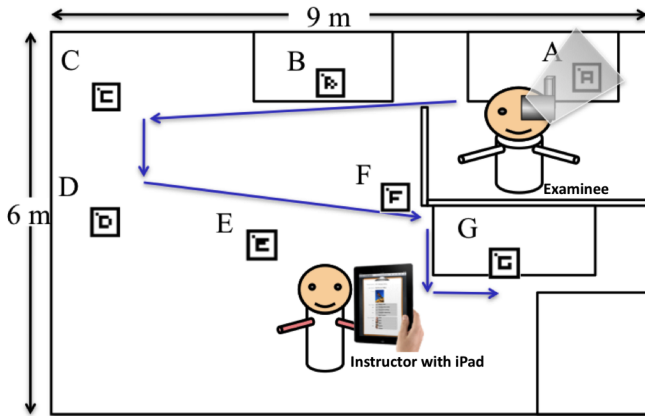


Figure 11: Evaluation of simulation disaster-relief training

Table 4: Average time required for instructors

(A-1) Time required to place AR markers	75.2 s
(A-2) Time required to place AR markers for additional scenario	18.4 s
(A-3) Time required to create scenario when sudden change occurs	7.6 s
(A-4) Time required to create additional scenario of removing injured people from scene	9.3 s

Table 5: Average time required for medics

(B-1) Triage time per person	32.6 s
(B-2) Time required to respond to notice new injured people	8.4 s
(B-3) Time required to take action in response to noticing sudden changes	47.1 s
(B-4) Time required to decide on which injured people to transport from scene	73.3 s

instructor place AR markers corresponding to the injured persons. The medic examines new injured people and performs triage (event B-2).

- At 6 min: The instructor generates a sudden change to injured person's condition (event A-3). The medic goes to the injured person and handles accordingly (event B-3).
- At 9 min: Injured people can be transported out from the scene (event A-4). The medic decides that two injured people assigned red tags should be transported (event B-4).

After the training, the instructor informs the medic of the elapsed time and the accuracy rate, and they exchange some comments. The examinees in this experiment were 10 pairs of students.

5.2.2 Result

Table 4 shows the average time required by the instructor for each event, and Table 5 shows the average time required by the medic for each event.

For instructors, the results show that the required time for each event was very short. It was possible to construct a training environment around one minute, and to modify scenario less than 20 seconds during training. On the other hand, the required time of medics needed to determine tag colors was an average of 32.6 seconds per injured person. Because triage is ideally performed in one minute per injured person or less, we can say this is a useful result. Events B-2, B-3 and B-4 show that medics completed correct actions in a short time when instructors added those events. These results show that medics responded quickly to various situations that occurred in triage. In addition, Table 6 shows the questionnaire results about the simulation training. In the evaluation, five is best, one is worst in each item. The questionnaire results show high score for all items pertaining to both medics and instructors. Thus we have confirmed the usability of the training using our system.

However, according to the network crossing and the light quantity in the training environment, our system was hard to detect of AR markers. Furthermore, the pronunciation that was hard to be performed voice recognition depending on a user was found. The improvement of the system which is not influenced by the personality of user and training environment will be necessary.

6 CONCLUSION

When a large number of people are injured at the same time, medics first need to perform triage in order to decide the priority for treatment depending on their severity of injury and the urgency of treatment. Nowadays, electronic triage systems that use PDA have been studied intensively. However, the operation of the devices may slow down the triage and become an obstacle for conducting necessary treatment in parallel.

In addition, disaster-relief training is very important for medics to be accustomed with real disaster situations and to know how the electronic triage system works before it is actually used. In the case of current training systems, biological information of injured people does not change as an actual disaster situation, and frequent training is impractical. Furthermore, there are many other problems in current training system.

In this study, we have developed an input interface of the first-triage where medics can get information such as injured vital signs through a monocular HMD and can control the system via voice input, so that medics' hands are free to conduct necessary immediate medical treatment even while doing triage. In addition, by using the first-triage input system with voice input, we have also developed disaster-relief training system with monocular HMD and AR markers. In our training system, injured people are represented by AR markers, so the preparation is easy. The presentation of injured person's information changes according to distance and direction of markers and their vital signs are generated at random by the system. In addition, the training result can be shared after the training.

After conducting the experiments of triage input and simulation disaster-relief training, we have confirmed that it is

Table 6: Results of training questionnaire

	Question	Score
Medic	Was it easy to understand the presentation of injured people information?	4.3
	Was it easy to see patient information on the HMD?	3.7
	Did voice recognition work satisfactorily?	3.1
	Was it easy to notice events?	4.6
	Did you see improvement in results?	4.1
	Did you work without feeling fatigue?	3.5
Instructor	Was it easy to understand the implementation status of triage?	4.3
	Was it easy to change the scenario?	4.5
	Did you feel like an active participant in training?	4.0
	Was it easy to find areas of improvement for the medic?	4.2
	Was it easy to implement training using this system?	4.1
	Did you work without feeling fatigue?	4.6

possible to determine the priority faster and more accurately with voice input than the current paper tag system, and make disaster-relief trainings more meaningful.

7 ACKNOWLEDGEMENT

This work is supported in part by a Grant-in-Aid for Scientific Research (B) No. 23300049 from the Ministry of Education, Culture, Sport, Science and Technology, Japan.

REFERENCES

- [1] "Simple Triage and Rapid Treatment," Disaster Medical Assistance Team, (<http://www.dmat.jp/index.html>)(in Japanese).
- [2] "The Triage Tag," Critical Illness and Trauma Foundation, Inc., (<http://citmt.org/Start/tag.htm>).
- [3] J. Takamatsu, M. Kishi, T. Ito, T. Nishimura, "Disaster Relief Activities in the Kansai Rosai Hospital after the Derailment Accident on the JR West Fukuchiyama Line," Japanese Journal of Disaster Medicine, vol. 13, no. 1, pp. 8–14, 2008(in Japanese).
- [4] "Triage Training," Fuji City, (<http://www.city.fuji.shizuoka.jp/hp/page000034100/hpg000034098.htm>)(in Japanese).
- [5] M. Numada, Y. Hada, M. Ohara and K. Meguro, "Development of IT Triage System (TRACY) to Share Regional Disaster Medical Information," 8CUEE Conference Proceedings, 8th International Conference on Urban Earthquake Engineering (Tokyo), Mar. 2011(in Japanese).
- [6] H. Matsuura, Y. Nakata, "Realistic and actual tabletop simulation is effective education for mass-casualty incident! - Disaster drill using Emergo Train System could improve the plans, the actions, and the inspection for mass-casualty management in the local reighting headquarters -," Japanese Journal of Disaster Medicine, vol. 17 no. 3, pp. 459–465, 2011(in Japanese).
- [7] A. Sonoda and S. Inoue, "Effective Use of Individual Information in the Emergency Activity," Proc. 18th Data Engineering Workshop 2007(in Japanese).
- [8] D. Malan, T. Fulfordjones, M. Welsh and S. Moulton, "Codeblue: An ad hoc sensor network infrastructure for emergency medical care," Proceedings of International Workshop on Wearable and Implantable Body sensor Networks, pp. 203–216, 2004.
- [9] J. G. Ko, J. H. Lim, Y. Chen, R. Musvaloiu-E, A. Terzis and G. M. Masson, T. Gao and W. Destler, L. Selavo and R. P. Dutton, "MEDiSN: Medical Emergency Detection in Sensor Networks," ACM Transactions on Embedded Computing Systems (TECS), 2010.
- [10] Y. Takahashi, K. Nagahashi, H. Kojima and K. Okada, "Proposal of Wound Person Information Management System using Second Triage at Disaster Scene," 78th Groupware and Network services, vol. 2011–GN–78 no. 4, pp. 1–8, Jan. 2011(in Japanese).
- [11] T. Gao, C. Pesto, L. Selavo, Y. Chen, J. G. Ko, J. H. Lim, A. Terzis, A. Watt, J. Jeng, B. rong Chen, K. Lorincz, and M. Welsh, "Wireless Medical Sensor Networks in Emergency Response: Implementation and Pilot Results," 2008 IEEE Conference on Technologies for Homeland Security, pp. 187–192, May 2008.
- [12] H. Kojima, Y. Takahashi, K. Okada, "Information Presentation System for Supporting Triage with START Method," Journal of Information Proceeding Society of Japan, vol. 53, no. 1, pp. 450–459, 2012(in Japanese).
- [13] N. Kaku, "Improvement of Medical Doctors, skills in Triage by Repeated Training," Japanese Journal of Disaster Medicine, vol. 7, no. 1, pp. 48–53, 2002(in Japanese).

A Study of Disaster Library System with a Field Agent to Learn a Sequence of Great Disasters

Taizo Miyachi^{*}, Gulbanu Buribayeva^{*}, Saiko Iga^{**}, and Takashi Furuhata^{***}

^{*} School of Science and Technology, Tokai University, Japan

^{**} Keio Research Institute of SFC, ^{***} University of Utah, USA
miyachi@keyaki.cc.u-tokai.ac.jp

Abstract Learning of safe evacuation from a sequence of great disasters such as earthquakes, tsunamis and accidents in nuclear power plants etc. and reconstruction of pleasant home town is very important. We propose a disaster library system “d-Library” with an seven layered core information base and a “field-agent” that allows citizens to not only study the right knowledge information but also get a chance to study how to ensure the safety with latest objective information in the world. We discuss methods to avoid danger by psychological causes such as catastrophe forgetting and normalcy bias. We also propose easy tools for learning the sequence of great disasters in order to easily evacuate in the safe direction against the great disasters and the concealment by the government and the companies.

Keywords: disaster library system, a sequence of disasters, learning evacuation, catastrophe forgetting, field-agent

1 INTRODUCTION

Unpredictable scale of disasters has caused tremendous number of deaths and serious man-made accidents in nuclear power plants [13]. People must quickly start evacuation and survive in case of surprise attack of a sequence of great disasters, such as earthquakes, fires, tsunamis [1], and heavy rain. Citizens should study the right knowledge information over many academic fields such as seismology, physics, tsunami science, psychology, geology, disaster science, safety engineering and information communication technology etc. in order to evacuate and reconstruct home towns with pleasant lifestyles. It is very difficult for the citizen to study the right knowledge and the latest information over many academic fields. Therefore education of the evacuation in a sequence of great disasters and man-made accident are very important for citizens to survive. We propose an easy disaster library system “d-Library” with field-agent and seven layered “core information” base over the many academic fields. d-Library with “field agent” not only provides the right knowledge and information but also

gives citizens a chance to check their safety in a sequence of great disasters. Citizens should avoid catastrophe forgetting (CF) [14] by a deluge of strongly impressed occurrences that are delivered by mass-media and survive by d-Library if the government and some companies would conceal inconvenient high risk information for them. We also propose easy tools for learning the sequence of great disasters in order to quickly evacuate in the safe direction and ensure the safety against the disasters, man-made accidents, CF and the concealment.

2 CORE KNOWLEDGE IN MANY ACADEMIC FIELDS AND HARD PREDICT OF DISASTERS

People that live in the Pacific Rim had to evacuate from the sudden attack of a sequence of great earthquakes, tsunamis and the accidents of Fukushima nuclear power plants in the East Japan Great Earthquake. All towns in Sanriku coast of 400 km in Japan were hit by great tsunamis and earthquakes. Human caused the serious accident in Fukushima nuclear power plants and radioactive contaminants by the explosions of plants contaminated both the terrestrial within 80km radius from the Fukushima plant and the ocean. Therefore citizens should study right “core knowledge” including scientific knowledge over many academic fields such as seismology, physics, tsunami science, psychology, geology, disaster, safety engineering and information communication technology etc. in order to safely evacuate and reconstruct home towns with pleasant lifestyle.

2.1 Hard predicts and concealing inconvenient information

Citizens should understand right core information and knowledge since major association of seismology in Japan (SSJ: The Seismological Association of Japan) is incompetent to predict when great earthquakes happen and



Fig.1 Great earthquake with fire, great tsunamis and an explosion of Fukushima III [10, 11]

both the Japanese Government at that time and “TEPCO (Tokyo Electric Power Co. Inc.)” concealed the serious accidents of Fukushima nuclear power plants. There are lots of problems for both safe evacuation and reconstruction of home towns with traditional lifestyle.

(a) Hard predicts and precursors of both earthquake and tsunami.

Japanese citizens should watch earthquake precursor site in the world and prepare the evacuation since SSJ could not predict not only “Hanshin-Awaji great earthquake Mj7.3” but also “Great East Japan Earthquake Mw9.0” at all. All Japanese should ensure that seismology can not predict great earthquakes, and should also find the good analysis of ancient documents that describe the histories of great earthquakes more than a thousand years. People should find precursors of great earthquakes by a high energetic ring of Lithosphere-Atmosphere-Ionosphere Coupling and strong low-frequency oscillation etc.

Predict of the height of tsunami is also very hard since the height depends on the terrain, such as estuaries and saw tooth coastline. The time rug of big tsunami is between a few minutes and several hours. Great tsunamis sometimes come without feeling earthquakes from a far away point. Results of analysis of the old documents show that a great tsunami occurred once in about a thousand year that is more than ten times of the human life span.

(b) Concealment of inconvenient information.

The Japanese Government at that time and TEPCO (Tokyo Denryoku (power) Co.) had concealed serious accidents in Fukushima nuclear power plants and put off the start of evacuation of residents.

Example 1. Long evacuation by bus and lots of death caused by the concealment of risk in nuclear power plants. Severe 34 patients of Futaba hospital and 98 people all residents of Long-Term Care Health Facility "Deauville Futaba" had to drive 230km in 14 hours to a high school gymnasium in Iwaki city on the third day although all 45000 citizens in Puripachi city in Russia were brought to a safe place for two hours forty minutes by 1100 large buses on the next day an accident in case of Chernobyl accident. In the end, 50 people in the 438 residents in both Futaba hospital and Deauville Futaba died in the moving. The major reason of the successful moving by the Russian Government was that the government had a set of rules of evacuation for the accident. The Japanese Government did not have such rules for the accidents because the Japanese Government created the myth “nuclear power plants are safe.”

(c) Ruining concealment.

Lots of children and young people were Atomic-bombed by strong radioactive contaminants since the government concealed a radioactive contaminant map that was urgently sent to the government from US embassy. The female victim would not be able to have a baby. “The long-term effects of "internal exposure" is also worried. The miserable life style of an Atomic-bomb victim might continue for several ten years. S/he feels sluggish and can not work well although the other persons can not watch the injured parts of

body and misunderstand that the victim would be lazy. “[2] Victims must acquire objective information like such map with deep meanings and the latest right information like prediction of height of tsunamis as “dynamic core knowledge” by mobile terminals and SNS in the Internet.

2.2 Normalcy bias and catastrophic forgetting

(1) Normalcy bias (NB). Normalcy bias is the “characteristic of the human mind” in risk psychology. A victim gathers convenient data for himself/herself and ignores inconvenient data for himself/herself in order not to panic. The victim thinks that I would be safe because the first tsunami warning is that the height of tsunami was 3 meters and the fence is taller than 3 meters. S/he repeatedly tells himself/herself “I am safe” and “This time is safe” although tsunami warning revised the height of the tsunami from 3 meters to 6 meters more. The victim with NB can not change his/her decision in serious situation in order not to panic. Citizens should mind that the first wrong predict of the height of tsunami causes serious damage.

(2) Catastrophe Forgetting (CF). Japanese mass-media continually reports a torrent of remarkable happenings such as large fires, explosions of gas tanks, broken buildings in cities, stopped train, accidents in highway tunnel, North Korean nuclear missile, avian influenza and tsunamis. There are also accidents such as fallen bookshelves and fallen products that were caused in the house by earthquakes. A citizen often forgets old occurrences and reduces his/her awareness against tsunamis since heads of the residents were occupied by a torrent of remarkable happenings, success stories to keep the safety against tsunamis for 37 years, and broken furniture by earthquakes in the house. Then frequent switch of such big topics in the wide range of different fields change citizen’s interesting into the newest topic and forced him/her to forget old but important occurrences. We call this “catastrophe forgetting (CF)” in disaster psychology.

Example 2. Forgetting the accidents in nuclear power plants. All Japanese were surprised at the scenes such that tsunami swallowed a whole town including lots of cars and all houses, the fire spread over a town, and several ten thousands of people at the stations that could not come home. Residents living near Fukushima nuclear power plants and Japanese citizens could not help reducing the awareness of the accidents in Fukushima nuclear power plants because of CF although a BBC news caster suspected that serious accidents might happen in the Fukushima plants and French government distributed a message in the homepage that French government prepared free airplanes for the French people staying within 80km radius from the Fukushima plant so as to evacuate to France.

(3) Bad environments by CF. Victims forgot the improvement of bad environments for weak persons and old patients in temporary shelters by CF after the Great East Japan Earthquake [11].

3.2 Seven layered information base and observation of high risk danger

The right knowledge information should be accumulated in seven layered information base (see Fig.3) since there are deluges of catastrophic occurrences in the internet and TV in Japan. Citizens should autonomously study not only strong points of a technology but also the weak points either in technology assessment (TA) or in the information base in order to start with the right understanding and reasonable consensus in the local area. Field-agent gives citizen a chance to acquire dynamic change of situations and to ensure the safety that citizen selected.

Wrong knowledge obstructed the important preparations of both back-up systems and emergency systems in the accidents of Fukushima nuclear power plants. TEPCO has been building dangerous facilities like radioactive water pools. TEPCO should open the safety of the facilities in the information base for citizens to check the safety and to find embedded risks by the assists in the world since technologies of nuclear power plant have many issues to be solved.

(a) Collective unconsciousness [6]. A person unconsciously takes an action according to “social custom in the community” based on a personal experiences and memories. Concrete conditions of release from the duties should be defined for the city worker such as volunteer firemen and social workers in the local society.

Example 5. Ms. Miki Endo working for Minamisanriku town kept announcing the tsunami warning for the residents without regard of her own death in the Great East Japan Earthquake and subsequent tsunami on 11th March 2011.

Example 6. Mr. Sato brought Chinese students to a safe hill and was killed by tsunami since he went downtown to rescue the weak persons again.

On the other hand side, citizens could not temporary move to the outside of home town since the other residents felt it treachery against the social custom.

(b) Tradition from old time and video. Citizens should utilize experiences that were handed from old time by traditions, and short message in geographical remains on a serious stone.

Example 7. Only a few people survived in Miyako city in case of Showa Sanriku Tsunami in 1933. They curved the

message “Never build a house lower than here” on the stone. Analyzing videos of great tsunamis clarifies the transition of strange phenomenon and the detailed causes of the danger.

(c) General Knowledge in home country. People should study how to start actions corresponding to the environments in emergency time by libraries, newspapers, TV and Web sites. General knowledge consists of major field: disaster, culture/social customs, experiences, map/mobility, industry, and education. Special education of a short term qualification for the reconstruction in stricken areas is indispensable for young residents to get jobs.

(d) Good ideas in other areas/foreign countries. Citizens should ask and share good ideas and back-up facilities from foreign countries in order to reduce a serious problem that nobody has experienced. d-Library should show the questions and introduce useful technologies to the public. Citizens should find advanced safe technologies in the world and compare it with the domestic solutions since even experts often could not know the cause of complex problem in advanced fields.

Example 8. The Swiss built both a hand-operated vent and a remover of radioactive contaminants for a Mark I type of atomic reactor in case of no power source about twenty years ago. Mr. Bruno Pillow thought that thinking loss of all power resource was a basic job in the 1980’s and recommended Japanese of TEPCO to build these facilities. TEPCO people never prepared those. This caused tremendous explosions of hydrogen gas in the Fukushima nuclear power plants accidents.

(e) Selected knowledge from SNS and the general public. Victims should ensure the safety during the evacuation by twitter since somebody could find the danger in latest information with photos etc. in the Internet. Public organizations could effectively solve problems in SNS in real-time and report the improved results.

Example 9. Notice danger. Lots of victims from Namie town could find strong radioactive contaminants in Iitate village where US embassy reported the danger by e-mail if they tweeted to ask their safety. Because many Internet users found the map of radioactive area in home page of US etc. [9]

Example 10. Unexpected problems of temporary shelters were discovered in Japan by twitter. The temporary shelters are not designed for a long time usage. “twitter” could

	[Field-agent]	
(g) Evolved knowledge	show	→ Solutions
(f) Serious problems + solutions	show	→ avoid CF
	remember	→ Problem + Solution
(e) Selected knowledge from SNS and the General Public	aware	→ ensure safety, by tweet/ FB
(d) Good ideas in other areas/foreign countries	show	→ Safe vent, Safe filter, Back-up facilities of power
(c) General knowledge information in home country	aware	→ Qualification of special car and machines
(b) Tradition from old time, Video	show	→ Monument, Phenomenon
(a) Collective Unconsciousness	free	→ Duty, NB, MB, PI, Social custom

Figure. 3 Eight layered knowledge information base of d-Library to observe Problem vs. solutions and history of statements

notice the managers of shelters and facilities the new problems.

(f) Serious problems + solutions. d-Library reserve (problem + solutions/ current situation) and statements that person in charge talked. Field agent shows citizens them in order to avoid CF and to accept both no excuse and no concealment.

(g) Evolved knowledge (EK). EK should be chosen from (a) to (f) and ranked based on the votes by local citizens. EK should be translated into easy explanation since even aged citizens should well understand EK from technology point of view, social ideology and custom point of view in their community with local conditions and instantaneously take the right actions to survive in emergency time.

Example 11. People that live either near a saw tooth coast line or on an active fault have the high risk of disaster. Anybody should easily access and study EK from various point views by a mobile terminal or an information distribution system.

4 PRIOR LEARNING BY AUTO-FLOW PICTURE BOOK AND PRACTICE OF CHECKING SAFETY

Prior learning, personal networks and a practice of checking safety are very important for citizens in order to avoid the misunderstanding and to quickly start effective actions for the safe survival against a sequence of great earthquake, fires, tsunamis, accidents in nuclear power plants, and concealment by companies and the governments. They are also indispensable for people without mobile terminal literacy. The personal networks that are built in daily life would provide the precious helps in the emergency time. However, it is difficult for the citizen to study many kinds of solutions since there are exact opposite actions between different disasters in the sequence of great disasters. We propose “Auto-flow Picture Book (APiBook)” (see Fig.4 and Fig.5) with automatic annotations for the prior learning and an application of practicing safety check in d-library so as to easily study the quick start of effective actions and to ensure the safe evacuation.

(1) Auto-flow Picture Book for Prior Learning

“APiBook” allows citizens to easily study knowledge and



Figure 4. A detailed information (a photo of Honolulu bay) in a display of picture book

information over many related academic fields for citizens by automatic annotation with pictures and quiz [8]. APiBook is especially indispensable for citizens without digital device literacy. The citizens need discriminately study many actions in an evacuation way for each disaster since there are exact opposite actions between the great disasters that subsequently happen. APiBook enables it by auto flow pictures with multimedia annotations. For example, APiBook shows that people should stay in a strong part of a house in case of earthquakes. They should go outside of the house and go up to a higher place in case of tsunamis. They should go to a strong building and stay in there in case of tornados. They should stay in the house at the explosions of nuclear power plants and quickly move to a safe place in the right direction.

Citizens can naturally understand the flow of actions with the condition while just watching the flow of pictures with annotations. We examined the digital picture book with four screens for about thirty people between teen's and 81 years old at Makiki church in Honolulu. We ensured that all users could easily use APiBook. Adults explained their children detailed information of Japanese American immigrants [5] that encountered a plantation life and a prison life caused by World War II. All people shared the space in front of APiBook with four screens and investigated on prior learning of experiences of their ancestors. Then they talked about their history and strengthened the bond between the ancestor and the descendants after the prior learning.

(2) Application of practicing safety check in d-library.

Objective analysis of the serious damages by great disasters should be supplied citizens since citizens can not afford to find the useful information in emergency time. They are hit by earthquakes and traffic jams etc. and the companies and the government conceal inconvenient facts and data for them. We propose “field agent” that shows awareness by the icon “check the safety” in order to easily practice safety check based on the latest information. One push of the icon allows a user to acquire objective information and many kinds of risk map like radioactive contaminant map of the Energy Department of State in US (see Fig. 7) from the outside of the interested parties or from foreign countries by one push of an icon (see Fig. 6).

Interview. “Can you avoid the risk by radioactive contaminant Map?”



Figure 5. Readable explanations in a picture book and basic operations

Subjects. A woman of 40's, a woman of 20's, two men of 30's, and two men of 60's

Answer. A woman of 40's answered that she could easily understand the danger of strong radiation by the map and quickly refuge to a safe place in the safe direction if the field agent would find the map in d-Library and show her it. Because Japanese usually watch the density map of pollen on TV or newspaper in spring and defends the pollen allergy by a special mask and special glasses. The other subjects answered as same as her answer and agreed with her answer.

Discussion. We could ensure that most Japanese could avoid the strong radiation by the map and the map was very useful since common Japanese citizens easily discover the danger in the maps. They could easily ensure his/her safety by the one push when they would worry their own situation. Here the most important preparations are that we should gather the excellent analyzers and watchmen all over the world in the Internet in order to find the indispensable knowledge and information from Big Data analysis. The daily cooperation between citizens in the internet is very important. The cooperation enhances the abilities of noticing strange changes and discovering embedded serious risks. We should prepare easy registration functions of new awareness and embedded risks for d-Library in order to expand the area of effective disaster cooperation and preliminary evacuation for the embedded risks.

The tsunami warning changed the prediction of the height of tsunami twice. The first revision was "from 3meters to 6meters" and the second revision was "from 6meters to more than 10meters." A citizen could also acquire awareness of the safety against giant tsunamis from field-agent in d-Library by only pushing the button of the check. The exercise of daily checking the safety by firemen and fishermen etc. is very important in this field since urgent check of the safety within several seconds after the revision would be very important in this field.

5. CONCLUSION

We described the importance of disaster library "d-Library" with field agent for citizens to autonomously grapple with embedded problems in both disasters and man-made accidents. The combination of (1) prior learning of scientific knowledge and ancestor's experience, (2) helps in

personal networks, and (3) assists by field agent in d-Library become useful in the sequence of disasters. The field agent gives citizens a chance so as to supplement additional solutions for new occurrences in collaboration through the Internet since even experts on each academic field could not predict both great disasters and the serious accidents in advanced system like nuclear power plants. We also discuss how to evacuate avoiding psychological weak points of human such as CF.

REFERENCES

- [1] Imamura, F., D. Subandono, G. Watson, A. Moore, T. Takahashi, H. Matsutomi, and R. Hidayat. Irian Jaya earthquake and tsunami cause serious damage, Eos Trans. AGU, 78(19), 197, doi:10.1029/97EO00128. (1997)
- [2] Nomura S. Mortality Risk amongst Nursing Home Residents Evacuated after the Fukushima Nuclear Accident, PLOS ONE website. (2013)
- [3] Fukushima Prefecture. What do you think of Japan's energy policy, Planning and Coordination unit of Fukushima Prefecture, pp.1-128. (2012)
- [4]hibi-zakkan.net. Of land for radioactive waste disposal site "Yaita, Tochigi Prefecture, Shiota" is discovered that it is just above the active fault, (2012) "http://hibi-zakkan.net/archives/16901126.html," "http://ameblo.jp/syuukitano/day-20120907.html."
- [5] Hojo R. Japanese Hawaii Immigrants History (Long rolled ehon). (2007-1990)
- [6] Jung C. Archetypes and the Collective Unconscious. (1959)
- [7] MIKAMI N. Challenges in the Practical Stage of Participatory Technology Assessment: From the Experience of GM Consensus Conference in Hokkaido, Japanese Journal of Science Communication, No.1. (2007)
- [8] Miyachi T. Active Learning of Multiple Cultures by Wide Rolled Ehon with Multiple 3D View Points, HICE2011. Parenti C. (2011).
- [9] U.S. Department of Energy. Aerial Monitoring Results Cumulative, "<http://energy.gov/downloads/radiation-monitoring-data-fukushima-area-32511>," March 25, (2011)
- [10] FNN archive. Never forget" Explosion and Fire in a

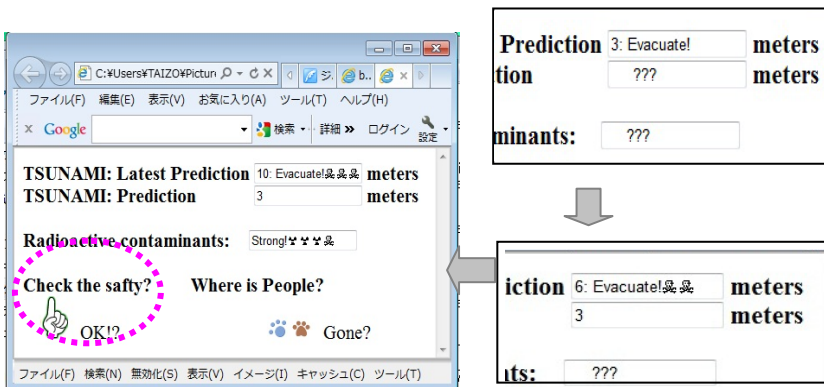


Fig. 6 One push check of the safety by "E-Safe" with two changes of the tsunami in tsunami warning

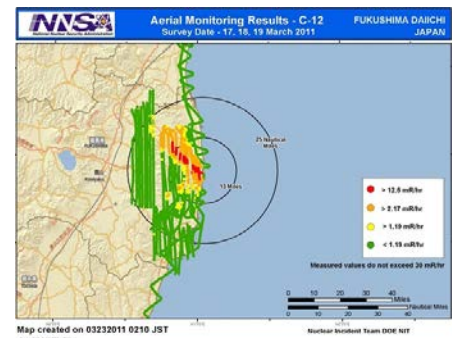


Fig7. The map of radioactive height of contaminants sent by US embassy

- complex in Chiba, “<http://www.youtube.com/watch?v=TaBx1j0hbHQ>,” Sept. 19, (2011)
- [11] TEPCO. Fukushima nuclear power plant III, <http://www.afpbb.com/article/disaster-accidents-crime/accidents/2846228/8206622?ref=ytopics>, March 21, (2011)
- [12] Wellman, Barry. The Community Question: The Intimate Networks of East Yorkers. *American Journal of Sociology* **84** (5): 1201–31. doi:10.1086/226906. (1979)
- [13] Parenti, Christian. Fukushima's Spent Fuel Rods Pose Grave Danger, *The Nation* March 15. (2011)
- [14] McCloskey, M. and Cohen, N., Catastrophic interference in connectionist networks: The sequential learning problem.” in G. H. Bower (ed.) *The Psychology of Learning and Motivation: Vol. 24*, 109-164, NY: Academic Press, 1989.
- [15] U.S. Department of Commerce, National Oceanic and Atmospheric Administration National Weather Service, Intergovernmental Oceanographic Commission, International Tsunami Information Center. Tsunami: the Great Waves,” <http://nws.noaa.gov/om/brochures/tsunami.htm>, (2013)
- [16] NOAA. National Weather Service, <http://nws.noaa.gov/om/brochures.shtml>, (2013)

A System to Help Creation of Original Recipes by Recommending Additional Foodstuffs and Reference Recipes

Mana Tanaka[†], Etsuko Inoue[‡], Takuya Yoshihiro^{‡*}, Masaru Nakagawa[‡]
[†]Graduate School of Systems Engineering, Wakayama University, Japan
[‡]Faculty of System Engineering, Wakayama University, Japan
^{*}tac@sys.wakayama-u.ac.jp

Abstract - Recently, many people use online recipe sites when they cook. As recipe sites are rapidly increasing in number, even recipe sites that have over 1 million appear. When the number of recipes is large, it is difficult for users to find the recipe that meets their requirements. To solve this problem, several web sites provide useful search interfaces, and several academic studies present methods to use the database effectively. We can classify these studies into two approaches, i.e., search for, and create recipes. In the studies to search for recipes, although these methods recommend recipes to users considering various aspect of users' requirements, users have no choice but to compromise because the recipe database would not include the one that perfectly satisfy the users' requirements. On the other hand, the study to create recipes only begins recently. Currently, many of them propose to recommend foodstuffs to add to, or delete from a base recipe. So, users who are beginners in cooking cannot cook the recommended foodstuffs because they have no idea how to cook them. In this study, we propose a method and a system to help users create their own original recipes. Specifically, the system first provides users with information that helps them to add or delete foodstuffs with a base recipes to determine the set of foodstuffs used in their original recipes. Next, it provides reference recipes, which is the selected recipes retrieved from the database, to help users get useful ideas on how to cook the set of foodstuffs determined in the previous step. We evaluated the system through a test experiment, and confirmed the effectiveness of the proposed system in creating users' own original recipes.

Keywords: Cooking, Recipes, Recommending Foodstuffs, Reference Recipes

1 Introduction

Recently, as reference materials for cooking, many recipe sites (see Table 1) have provided various recipes in the Internet, and are populated instead of the traditionally used materials such as books, magazines, and TV programs on cooking. Having grown rapidly, these recipe sites include a significant number of cooking recipes. For example, Cookpad [1], one of the most famous Japanese recipe sites, includes more than 1 million cooking recipes. However, among these tremendous number of recipes, it is difficult to find recipes for people that meet their requirements. To solve the difficulty, not only the recipe sites prepare keyword or categorical search in their sites, but also many academic studies have been performed that tries to enrich users' experiments to satisfy users' requirements[6]. These studies are classified to two approaches: one

Table 1: Recipe Databases

Site name	# of recipes	URL
Cookpad[1]	140 million	http://cookpad.com/
CDkitchen[2]	120 thousands	http://www.cdkitchen.com/
RecipeSource[3]	70 thousands	http://www.recipe-source.com/
All recipes[4]	50 thousands	http://allrecipes.com/
E-recipe[5]	20 thousands	http://erecipe.woman.excite.co.jp/

is to “search” recipes, and the other is to “create” recipes.

The approach that “search” recipes typically computes the similarities between recipes from various aspects to find or recommend recipes that is likely to meet users' requirements. This approach enables us to retrieve recipes according to a common characteristics between recipes that reflect requirements or testes of users. However, these methods only search for recipes from a limited set of recipes in a database so that users will possibly find recipes that nearly meets their requirements, but cannot do the perfect one that completely satisfies their requirements.

On the other hand, the approach that “create” recipes is potentially able to retrieve the perfect recipes that meets the requirements, although they needs higher level techniques. Towards this goal, several proposals are presented that help users create their own recipes by modifying existing cooking recipes. For example, there is a study that recommends foodstuffs that are likely to be added/deleted in a recipe easily [10]. However, because they do not support to design the operational steps to cook dishes using the modified set of foodstuffs, it is difficult for many beginners to complete the original recipes, i.e., they do not know how to cook the modified foodstuffs. We have to help constructing the operational steps of cooking in their original recipes.

In this paper, we present a method and a system that not only recommend foodstuffs that can be added/deleted in a recipe, but also present information to help users constructing the operational steps of their own original recipes. Specifically, in recommending foodstuffs, we display several sorts of supporting information that indicates the foodstuffs likely to be added/deleted in the base recipe from several aspects. In helping operational-steps design, we present reference recipes that includes the operation steps to cook the added foodstuffs in combination with the other foodstuffs in the base recipe. With these two functions, we help users create their own original recipes.

The remainder of this paper is organized as follows: In Section 2, we present the related work. In Section 3, we present the design of our system to help users create their own original recipes, and in Section 4 we give the algorithms and formula

that underlie in our system. We evaluated the effectiveness of our study in Section 5, and finally we conclude the work in Section 6.

2 Related Work

2.1 The Current State on Recipe Sites

In the last decade, people have become to use recipe sites as materials to refer cooking recipes, in addition to books, magazines, and TV programs. These web sites are strongly supported not only by young people but also by middle-aged people, mainly because (1) they are free to access, (2) they include significant number of recipes, and (3) they are available at all time.

In February 2009, iShare Corporation surveyed the media to refer when people cook in Japan [7]. 274 people answered for it, in which 53.3% is men and 46.7% is women, including 11.6% of twenties, 46.3% of thirties, 31.6% of forties, and 10.5% of others. The result shows that people use recipe web sites the most when they cook. Specifically, 58.8% of men and more than half of the twenties and forties (men and women) answered to use web sites the most among other media.

The most outstanding reason that people refer to these recipe sites is the number of recipes available in these site, which is far larger than those of books or magazines. In general, users has various and sensitive requirements for recipes such as likes and dislikes of foodstuffs, allergies, foodstuffs stored in refrigerator (so they want use them), etc. To find recipes that satisfy these sensitive requirements, recipe sites where a vast number of recipes are stored are convenient.

In the recipe sites, almost all sites provide keyword-search function, with which users try to search for recipes that meet their sensitive requirements. Keyword search alone, however, hardly enables users to find the best recipe that meets users' requirements, since the number of search results is also large in proportion to the total number of recipes. For example, if we search for recipes with two keywords "potatoes" and "gratin" in a major recipe site "cookpad," in which 1.4 million recipes are stored in it, we will have about 4,000 recipes retrieved from the database. Consider that an user checks the results one by one to find recipes that satisfy his/her own requirements the most. It is hard and laborious to find suitable recipes with keyword search alone.

Many sites provide further functions to help users to find suitable recipes effectively. Specifically, many sites provides a function that narrow the search space using several properties such as objective (e.g., for health, for beauty, for body building), categories (e.g., main course, soup), cooking time, and cooking methods. Each sites provides various original filters to help searching. As shown above, user interfaces to support efficient recipe searching have been developed with significant care.

It is, however, still hard to find recipes that completely satisfies users' sensitive requirements, as long as we tries searching for recipes from a limited set of recipes. Customising recipes is essential to achieve higher level recipe searvices on the web.

2.2 Related Work

Studies that target recipe data are roughly classified into two categories: studies that search for recipes, and those create recipes. In this section, we describe the state of the art of these two categories of studies.

Among the studies that search for recipes, we introduce several searching methods based on foodstuffs, as the related work with this paper. Ueda et al.[8] proposed a method to recommend recipes that considers likes and dislikes of users. Their method recommends recipes based on the frequency of foodstuffs appearing in an user's cooking history. Iwagami et al.[9] proposed a method to recommend recipes that also considers likes and dislikes of users. Their method first acquires foodstuffs that the user likes to use from the history of recipes users refer in the system, and second computes scores of recipes according to the acquired information of each foodstuff. These two proposals to recommend recipes help efficient searching for recipes that considers users' tastes, by utilizing additional data that include the history of recipes referred in the system.

On the other hand, the approach that creates recipes is still in the stage of beginning. We can only introduce several studies that treat addition and deletion of foodstuffs in a recipe. Shidochi et al.[10] proposed a method that suggests foodstuffs to add to the base recipe. Their method is based on a typical pattern of cooking steps in each sort of dishes; they suggest an alternative foodstuff that can be used in each operational step of cooking, under the assumption that the sort of dishes and so the typical pattern in cooking is given. Tsukuda et al.[11] also proposed a method to suggest foodstuffs to add to, or delete from the input recipe, based on the combination of foodstuffs included in the set of recipes in the database. They compute the stability of a combinations of foodstuffs from the frequency of combinations that appear in the recipes in the database. Based on this stability scores, their method suggest a foodstuff to add that increases the stability of the set of foodstuffs. Although these studies suggest foodstuffs to add to, or delete from the input recipe, they do not consider the operational steps of cooking in the newly created original recipes. Thus, unless the users are accustomed to cook, it would be difficult to complete the recipe with the new set of foodstuffs by determining the operational steps to complete their original recipes.

3 System Design

3.1 Requirements

As users of the proposed system, we suppose people who are not so accustomed to cooking that they cannot modify recipes by themselves, although they can cook by pursuing a recipe that includes full description of operational steps. In fact, this kind of people frequently refers recipe sites when they try cooking. However, as mentioned in the previous section, it is laborious for them to search for suitable recipes that meets their sensitive requirements.

In this case, many people will come to the idea to modify the existing recipes to meet their own requirements. We de-

signed our system to be useful in this case. In the case, users first try to select a base recipe, and try to change foodstuffs to meet the requirements they have. Users then find a problem that they do not know which foodstuff is suitable to add/delete in combination with other foodstuffs in the base recipe. To help users on this point is the first task for us to perform.

When a set of foodstuffs is determined, users next try to decide how to cook them. However, because the cooking operations are different among foodstuffs (i.e., imagine that some foodstuffs such as fish require special handling or preliminary operations, or even for basic foodstuffs, cutting size may be different for each foodstuff), it is hard to find the appropriate operation steps for the modified set of foodstuffs. To help users on this point is the second task for us to perform.

In this paper, we design an information system that helps users on these two points, so that users can create their original recipes by modifying the base recipe to meet their sensitive requirements, using helpful informations suggested by the system.

3.2 Functional Design

We designed a system that provides valuable information for users to help them in the two troublesome situations in creating recipes, i.e., (i) the first is the situation where users select foodstuffs to add/delete, and (ii) the second is the one where users decide cooking operations over the selected set of foodstuffs. In this section, we describe a basic design of the functions in our system for these two target situations.

We first describe the functions to support users selecting foodstuffs to add/delete. When people cook, they easily think of several requirements such as: “I want to eat omuraisu,” “I want to eat salmon in today’s dinner,” or “I want to use spinach left in my refrigerator.” These examples imply that users can easily think of dishes (e.g., omuraisu) they want to have, and also of the foodstuffs they want to use; they can easily select the sort of dishes and foodstuffs to use. It is difficult, however, for users to judge whether a set of foodstuffs goes well in cooking. Note that there are combinations of foodstuffs that are easy or difficult to cook well together. So, it is valuable that our system provides information that helps users to be aware of good and bad combinations, for each sort of dishes.

In our design, we implement a function that provides three sorts of information that helps users select foodstuffs to add/delete, as follows:

- (a) Frequency of foodstuffs used in each sort of dishes.
- (b) The degree of compatibility between a newly added foodstuff and the other foodstuffs.
- (c) Foodstuffs that have good compatibility with a current set of foodstuffs.

First, (a) provides information that indicates which foodstuffs are used frequently in the sort of dishes chosen by users. For each sort of dishes (e.g., omuraisu), the foodstuffs used frequently are considered compatible, and are suitable to use

in cooking it. This information is considered useful to know the generally used foodstuffs for each sort of dishes.

Second, (b) provides information that indicates the degree of compatibility between a newly added foodstuff and the other foodstuffs included in the editing recipe. When a user add a new foodstuff in our system, we display the compatibility between the new foodstuff and the other foodstuffs listed in the screen. This function is useful for users to know which foodstuff is suitable for the current set of foodstuffs through trial and error of repeated addition and deletion of foodstuffs. Also, it is useful to decide foodstuffs to delete, instead of the newly added foodstuffs.

Third, (c) provides information that indicates the foodstuff that have good compatibility with a current set of foodstuffs. This function recommends foodstuffs to users to add into the current set of foodstuffs.

These three sorts of information are displayed in the user interface of our system with small icons beside the name of foodstuffs. Because users see these icons easily, our system enables users to operate it intuitively. The detail of the user interface is shown in Section 3.3.

As the functions that help users to decide cooking operations over the selected set of foodstuffs, we display recipes retrieved from the database that in high probability includes the cooking operations over the added foodstuff. Hereafter, we call such recipes as *reference recipes*. The reference recipes that the system displays should involve a similar set of foodstuffs, and simultaneously should have similar pattern of cooking operations with the base recipe.

The reference recipes are displayed for users in the order of the similarity between the set of foodstuffs selected by users and the set of foodstuffs in each reference recipe. By displaying several reference recipes, users can refer variety of cooking operations for the newly added foodstuffs, which we expect users to refer to for an idea how to cook them in their original recipes. Also, we can expect an effect that the system suggests users not only conventional cooking operations, but also rare and surprising operations, which give users a precious hint for their own original dishes.

3.3 User Interfaces

Based on the functional design described in the previous section, we designed the user interface of our system. In this section, we will show how users use our system by introducing our interface design.

The overview of the usage of our system is shown in Fig. 1. Users use the system with the following three steps. First, users select a sort of dishes to cook among several candidate sorts of dishes provided by the system (①). Hereafter, we call the typical recipe of the selected sort of dishes a *base recipe*. Second, users determine a set of foodstuffs to use in their original recipes by adding (or deleting) foodstuffs to (or from) the base recipe (②). Third, users decide how to cook the set of foodstuffs determined in step ② with the help of reference recipes provided by the system (③). In the following, we explain the detail of steps ② and ③.

In step ②, users determine a set of foodstuffs by modifying the base recipe selected in step ①. The user interface of this

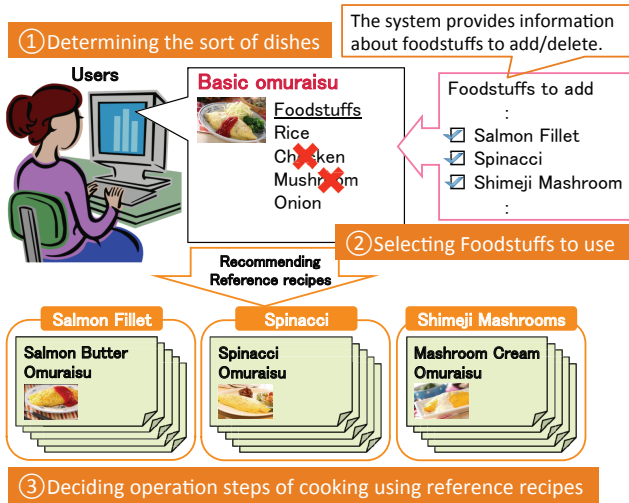


Figure 1: Overview of Usage of Our System



Figure 2: User Interface of the System

step is shown in Fig. 2. In (A), the base recipe selected in step ① is displayed. In (B), users add or delete foodstuffs for their original recipe. In this field (B), many candidate foodstuffs to add or delete are listed, which includes all the foodstuffs in the base recipe.

Because the number of all existing foodstuffs is tremendous, we have to create a subset of them to be displayed in this field. We limited the number of foodstuffs to display according to the criterion (a), the frequency of each foodstuff used in all the recipes that belong to the sort of dishes selected in step ①. The criterion (a) is introduced in Section 3.2, and is described specifically in Section 4. Also, to have users select foodstuffs to add/delete easily, we classified foodstuffs into several categories according to “the standard tables of food composition” [12]. In the user interface of our system, each category of foodstuffs (e.g., vegetables, meat, fish, mushrooms, etc.) forms an “island” with its own color, as shown in Fig. 2.

In this field (B), users add or delete foodstuffs repeatedly by clicking icons placed at the rightside of the text. (The zoomed image of Field (B) is shown in Fig. 3.) For each operation of users, the system reacts with the three types of information (a)(b) and (c) recalculated and displayed at the rightside of the foodstuff texts.



Figure 3: User Interface (Zoomed)

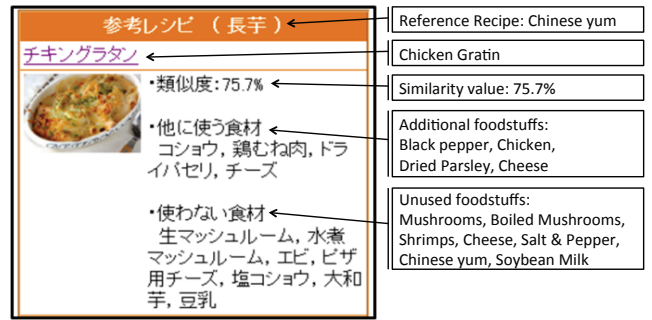


Figure 4: Layout of a Reference Recipe Displayed in the field C of Fig.2

The first type of information, i.e., (a) the frequency of each foodstuff used in the sort of dishes, is expressed by the text size of each foodstuff. Fig. 3 is the zoomed snapshot of field (B). Frequently used foodstuffs are shown with bigger fonts, whereas rarely used foodstuffs are shown with smaller fonts. Here, the numbers at rightside of foodstuff texts in brackets are the number of recipes that includes the foodstuff.

The second type of information, i.e., (b) the degree of compatibility between a newly added foodstuff and the other foodstuffs, is expressed by icons. The icons are displayed at the rightside of foodstuff texts when a user newly adds a foodstuff, where the different icons are displayed according to the level of compatibility between them, as shown in Fig. 3. If a foodstuff is very compatible with the new one, a star icon appears. If a foodstuff is moderately compatible, a red arrow with upper direction appears. If a foodstuff is not compatible at all, no icon appears. And, if a foodstuff and the new one are in bad combination, a blue arrow with lower direction appears.

The third type of information, i.e., (c) foodstuffs that have good compatibility with a current set of foodstuffs, is expressed with “heart” icons, displayed in the rightside of the foodstuff texts. The arithmetic formula to compute the degree of three criteria (a)(b) and (c) is presented in Section 4.

Next, in the step ③, users decide how to cook the set of foodstuffs determined in step ② with the reference recipes provided by the system. When users add a foodstuff, the system searches the database for the recipes that can be used for reference in deciding the cooking operations, and the found reference recipes are displayed in the field (C). As the reference recipes, the system selects the ones that include the added foodstuffs, and simultaneously that the similarity of the set of foodstuffs to those selected in step ② is high. Fig. 4

Table 2: Example of Reference Recipes Listed in the System

Chinese Yum	White Leek	Soybean Milk
Chicken gratin	Cod gratin	Shrinsps and macaroni gratin
Chinese yum and shrimp gratin	Chinese yum gratin	Potatoes and shrimp gratin
Seafood and chinese yum gratin Japanese style	Oysters and trout gratin	boiled eggs and soybean milk gratin
Chinese yum gratin	Chinese yum gratin Japanese style	Seafood and chinese yum gratin Japanese style
Chinese yum gratin Japanese style	Chinese yum gratin Japanese style	Mushroom gratin

shows the example of the layout of reference recipes displayed in the field ©. In the field, there are several items of useful information such as the similarity value, the additional foodstuffs (i.e., the foodstuffs used in the reference recipe only), and the unused foodstuffs (i.e., those used in the original recipe only). When users click the field, users can refer full information of the recipe in another window.

Table 2 shows the examples of the listed reference recipes for each of added foodstuffs “chinese yum,” “white leeks,” and “soybean milk,” in the case where the base recipe is “gratin”

4 Underlying Algorithms

4.1 An Algorithm to Recommend Foodstuffs

In this section, we introduce an arithmetic criteria to provide information to help users in step ②. We first present the criterion (a), the frequency of each foodstuff used in all recipes that belong to a sort of dishes. We define the set of recipes R in the database as

$$R = (r_1, r_2, \dots, r_n), \quad (1)$$

where $r_k (1 \leq k \leq n)$ is a recipe and n is the number of recipes in the database. We also define a set of foodstuffs M . If a foodstuff $m \in M$ is included in a recipe r_i , we write $m \in r_i$. Let C denote a sort of dishes selected in the step ①. Then, the frequency $F_C(m)$ of a foodstuff m in a sort of dishes C is represented as follows:

$$F_C(m) = \frac{|\{r | r \in C \text{ and } m \in r\}|}{|\{r | r \in C\}|}. \quad (2)$$

In our system, we use four sorts of fonts according to this value $F_C(m)$.

Next, as for the criterion (b), the degree of compatibility between a newly added foodstuff and the other foodstuffs, we let the degree of compatibility among two foodstuffs m_i and m_j be $Comp_C(m_i, m_j)$, as follows:

$$Comp_C(m_i, m_j) = \frac{|\{r | r \in C \text{ and } m_i \in r \text{ and } m_j \in r\}|}{|\{r | r \in C \text{ and } (m_i \in r \text{ or } m_j \in r)\}|}. \quad (3)$$

In our system, we classify the combinations of foodstuffs m_i and m_j with the value $Comp_C(m_i, m_j)$ into four classes, and display icons accordingly.

Finally, we present the criterion (c), foodstuffs that have good compatibility with a current set of foodstuffs. We let r_0 be the original recipe created by a user, and $S(r_0, r_i)$ be the *similarity* between these two recipes. Then, $S(r_0, r_i)$ is

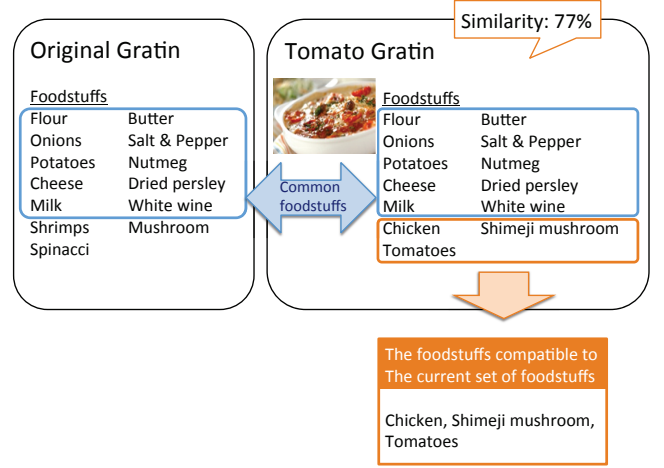


Figure 5: Computing Foodstuffs Compatible with The Original Recipe (Criterion (c))

computed according to the ratio of commonly used foodstuffs, as follows:

$$S(r_0, r_i) = \frac{|\{m | m \in r_0 \text{ and } m \in r_i\}|}{|\{m | m \in r_0 \text{ or } m \in r_i\}|}. \quad (4)$$

In our system, we retrieve a set of recipes from the database with all recipes r_i that satisfies $S(r_0, r_i) > 0.7$, and display the icon to the foodstuffs included in the retrieved recipe set, except for the foodstuffs included in r_0 (Fig. 5).

Note that the computational complexity for each criterion (a)(b) and (c) is $O(|R|)$, where $|R|$ is the number of recipes in the database.

4.2 An Algorithm to Select Reference Recipes

A list of reference recipes are displayed when users add a foodstuff when they are creating their recipes. Thus, the reference recipes are computed from the added foodstuff, the sort of dishes, and the current set of foodstuffs in the original recipe. Specifically, the system first retrieves all recipes that belong to the sort of dishes, and that include the added foodstuff. The system then sorts the retrieved recipes by the similarity between each of them (r_i) and the original recipe (r_0) presented as formula (4). Top 5 of the recipes are displayed as the reference recipes.

The time complexity for searching reference recipes is the same as the algorithms presented in the previous section. That is, we can compute all the criteria (i.e., (a)(b) and (c)) and the reference recipes within a single scan for each recipe.

Table 3: Foodstuffs Specified to Use in the Experiment for each sort of dishes

Gratin	Omuraisu
Chinese yum	White radish sprout
White miso	White leek
Mince meat	Ginger

5 Evaluation

5.1 Methods

To evaluate the effectiveness of the proposed system in creating original recipes, we conducted an experiment. In the experiment, we asked users to create their own original recipes, i.e., to write down them, using the proposed system with the following two conditions, and compared the results.

Condition 1: Using the proposed system with all functions.

Condition 2: Using the proposed system with all functions, except for reference recipes.

Because what is new in this system in the literature is the function to provide reference recipes, we confirm the effectiveness of the reference recipes in creating original recipes by comparing the systems with and without reference recipes. Note that, to conduct fair comparison, when users use the system without reference recipes (i.e., the case of Condition 2), we allow users to search recipes using the web site E-recipe[5] to decide their cooking steps in their own original recipes. The sorts of dishes we asked users to create recipes were “gratin” and “omuraisu.”

In our system, we imported the recipes from the web site E-recipe[5]. We select the site E-recipe because the recipes in this site relatively includes low fluctuation of words and the name of foodstuffs. Although the fluctuation is relatively low, we integrated names of foodstuffs according to the reference [12]. (E.g., if both “potato” and “danshaku” are used in the recipes, we integrate them into a word “potato.”)

The experiment is done as follows: We form two groups of users A and B. For users in group A, we first ask them to create their own original recipes with the base recipe “gratin” in Condition 1, and next asked them to do the same operation with the base recipe “omuraisu” in Condition 2. For users in group B, we asked them to do the same with the exchanged base recipes, i.e., they first create recipes of “omuraisu” in Condition 1, and next “gratin” in Condition 2. We also asked them to write down the operational steps of their own original recipes concisely and to answer the questionnaire when they finished creating each of their original recipes.

In the experiment, we specified foodstuffs that users must use in their own original recipes, as a “requirements” in their modification of recipes. Namely, for each of two sort of dishes, we specified three foodstuffs to use, while other foodstuffs are free to use. The three foodstuffs specified are shown in Table 3. We selected these three foodstuffs because they are not frequently used in the given sort of dishes, and also because they have more than one ways in cooking. By specifying foodstuffs that have several ways to cook, we intend

to have users being not easily able to decide the operational steps to cook their original recipes.

In the questionnaire, we have questions on how the three sorts of information for steps ② is useful in selecting the foodstuffs in their recipes. Also, we have questions on how the reference recipes for steps ③ are useful in deciding operational steps in cooking. For each questions, users answer with a 5-grade rating, where 5 means “very useful” or “strongly agree,” and 1 means “not useful at all” or “strongly disagree.” Furthermore, we checked that the written operational steps in their original recipes are proper or not, and compared them between the cases with and without reference recipes.

5.2 Results

In Table 4, we show the results on usefulness of the three sorts of information for adding/deleting foodstuffs. In the results of questions (i), (ii), and (iii), all medians and modes are equal to or more than 4, meaning that users answered that these three sorts of information were useful in selecting foodstuffs to add/delete. However, for the question (iv), users answered that they did not feel like sufficiently easy to select foodstuffs to add/delete with this system. One of the possibility that the results indicate is that, to select foodstuffs to add/delete in creating recipes, users may require not only the information on compatibility among foodstuffs, but also the information that recall the idea of creating recipes.

In Table 5, we show the results on usefulness of reference recipes. In the results of questions (v)-(ix), all medians and modes are equal to or more than 4 with reference recipes, whereas they are equal to or lower than 3 without reference recipes. There were big difference between the cases with and without reference recipes. The difference was confirmed by checking p-values in t-test of the two cases. The results are shown in Table 6, where the statistical significance was confirmed in all the questions (v)-(ix). We also had a result that most users answered that the reference recipes are better to be recommended automatically. Consequently, we concluded that the reference recipes are useful in deciding how to cook foodstuffs in their own original recipes.

On the other hand, as a result of checking the recipes written by users, we found that the recipes created with reference recipes are all proper, i.e., it does not include wrong operations, whereas those without reference recipes includes several faults. For example, in cooking gratin, some original recipes cut raw chinese yum and throw directly into white source. Note that chinese yum are usually lightly fried before mixed with white source. As another example, an original recipe first bakes the gratin in an oven, and after that, puts the fried ingredients on it. The ingredients are usually mixed with white source before the gratin with white source is baked. The reason why users made such mistakes would be that they referred the recipes that belong to other sorts of dishes. Consequently, they only understood how to cook it, but not understood the timing and the sequence of operations in gratin. This also indicates that the reference recipes recommended from the same sort of dishes work effectively to decide operational steps in cooking in their own original recipes.

Table 4: Results: usefulness of three sorts of information for adding/deleting foodstuffs

Questions		Evaluation					Median	Mode
		5	4	3	2	1		
(i)	The information “frequency in the sort of dishes” was useful.	7	5	4	4	0	4	5
(ii)	The information “the degree of compatibility between a newly added foodstuff and the other foodstuffs” was useful.	5	13	1	1	0	4	4
(iii)	The information “foodstuffs that have good compatibility with a current set of foodstuffs” was useful.	4	14	1	0	1	4	4
(iv)	Selecting foodstuffs to add/delete was easily done.	2	7	5	6	2	3	4

Table 5: Results: usefulness of reference recipes

Questions		Evaluation					Median	Mode
		5	4	3	2	1		
With reference recipes	(v) I am satisfied with my original recipe.	6	8	5	1	0	4	4
	(vi) Deciding how to cook the selected foodstuffs was easily done.	8	8	4	0	0	4	5
	(vii) This system is useful to get an idea in creating recipes.	10	9	1	0	0	4.5	5
	(viii) I would like to use this system again.	4	15	1	0	0	4	4
Without reference recipes	(v) I am satisfied with my original recipe.	0	6	7	6	1	3	3
	(vi) Deciding how to cook the selected foodstuffs was easily done.	0	3	4	9	4	2	2
	(vii) This system is useful to get an idea in creating recipes.	1	5	6	7	1	3	2
	(viii) I would like to use this system again.	0	3	6	10	1	2	2
(ix)	It is better that the reference recipes are automatically recommended.	18	2	0	0	0	5	5

Table 6: Results: usefulness of reference recipes (p-values)

Questions	With reference recipes		Without reference recipe		p-value
	Average	Stddev	Average	Stddev	
(v) I am satisfied with my original recipe.	3.95	0.89	2.90	0.91	0.0003
(vi) Deciding how to cook the selected foodstuffs was easily done.	4.20	0.77	2.30	0.98	0.00000003
(vii) This system is useful to get an idea in creating recipes.	4.45	0.92	2.90	1.02	0.000001
(viii) I would like to use this system again.	4.15	0.60	2.55	0.83	0.00000001

6 Conclusion

In this paper, we proposed a method and a system to help users to create original recipes. The proposed system provides users with the information that helps users to select foodstuffs to add to, or delete from their own original recipes, and also with reference recipes that helps users to decide operational steps in cooking in their original recipes. With this system, users are able to create their original recipes that meet their own requirements with the helpful computational aids.

We evaluated the system how effectively it helps users to create their own original recipes. Through the experiment to create original recipes, we confirmed that the proposed system is useful in creating their own recipes, and this proves the effectiveness of the system.

One of the challenges for the future is to recommend reference recipes for various objectives of users, e.g., reference recipes for basic cooking methods, or those for stimulative idea in cooking, etc. Other customizations and characterizations to fit the system to users' various requirements would also be a possible task for the future.

REFERENCES

- [1] Cookpad, <http://cookpad.com> .
- [2] CDkitchen, <http://www.cdkitchen.com/> .
- [3] RecipeSource, <http://www.recipesource.com/> .
- [4] All recipes, <http://allrecipes.com/> .
- [5] E-recipe, <http://erecipe.woman.excite.co.jp/> .
- [6] Kiyoharu Aizawa and Ichiro Ide, “Foods and Computing,” IPSJ Magazine, Vol.52, No.11, pp.1368-1408, 2011 (In Japanese).
- [7] iShare corporation, A Survey on the source of recipes to refer, <http://release.center.jp/2009/03/0602.html> (In Japanese).
- [8] M. Ueda, K. Ishihara, Y. Hirano, S. kajita, and K. Mase, “Recipe Recommendation Method Based on Personal Use History of Foodstuff to Reflect Personal Preference,” DBSJ Letters, Vol.6, No.4, 2008 (In Japanese).
- [9] M. Iwakami and T. Itou, “An Implementation of a Recipe Recommendation System with User’s Preference Order,” The 24th Annual Conference of the Japanese Society for Artificial Intelligence, 2010 (In Japanese).
- [10] Y. Shidochi, T. Takahashi, I. Ide, and H. Murase, “Discovery of Replacable Materials from Cooking Recipe Texts,” The 22th Annual Conference of the Japanese Society for Artificial Intelligence, 2008 (In Japanese).
- [11] K. Tsukada, S. Nakamura, T. Yamamoto, and K. Tanaka, “Recommendation of Addition and Deletion Ingredients Based on the Recipe Structure and Its Stability for Exploration of Recipes,” IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences J94-A(7), pp.476-487, 2011 (In Japanese).
- [12] Ministry of Education, Culture, Sports, Science, and Technology, Japan, “Standard Tables of Food Composition In Japan Fifth Revised and Enlarged Edition,” 2005 (In Japanese).

An Estimation Method of Calorie Consumption in Activities of Daily Living based on METs Values

Yoshitaka Nakamura[†], Yoshiki Matsubayashi[‡], Yoh Shiraishi[†], and Osamu Takahashi[†]

[†]School of Systems Information Science, Future University Hakodate, Japan

[‡]Graduate school of Systems Information Science, Future University Hakodate, Japan
{y-nakamr, siraisi, osamu}@fun.ac.jp

Abstract - It is important to know the calorie consumption at the time of life activities for prevention of Lifestyle diseases. In this paper, we propose a method of measuring calorie consumption with high precision. Although there are some studies to calculate the calorie consumption by METs values adapting the state estimation method of the person using an accelerometer, it is insufficient to calculate calorie consumption with high precision for the estimation of the states such as road condition and moving speed. The proposed method aims to calculate calorie consumption with high precision during the life activity by estimating road condition and moving speed in addition to the items of conventional estimation using the accelerometer of smartphone. Finally, from the accuracy evaluation of the proposed method, estimation errors were able to be reduced by about 83% from the conventional methods.

Keywords: Provide up to five keywords to be used for future on-line publication searches and indexing.

1 INTRODUCTION

Lifestyle diseases occupy the biggest factor of cause of death in Japan now [1]. For the prevention of these lifestyle diseases, it is important to balance calorie intake and calorie consumption. Recently, the calorie intake comes to be recorded in many food and be easily known, but calorie consumption is difficult to recognize. Therefore the method to easily calculate calorie consumption is required.

The METs values [4] is known as the method to calculate calorie consumption easily from the strength and the duration of the activity depending on the types of life activities. By this method, we can calculate calorie consumption if we can know what kind of activity a person takes and how long the person continued the activity. In various kinds of activities, activities in daily living have a big influence on calorie consumption of the day, because the activities occupy most of one-day activities [2]. Therefore, we can calculate calorie consumption if we can classify the state of daily living activities precisely. But it is necessary to observe the continuous state of activities to classify the state of daily living activities.

On the other hand, with the development of the mobile terminals such as the smartphones, mobile terminals come to equip with many sensors which can measure various phenomena precisely. A kind of sensor such as the acceleration sensor can sense motion of human beings and the various kinds of sensors can record as data sequences. In addition, these terminals are convenient for recording daily living activities, because the terminal tends to be always worn in daily living

activities. Using these sensors equipped with by such terminal, there are several methods to estimate daily activities [3] and method to calculate calorie consumptions [5], [6]. However, though calorie consumption greatly varies with moving speed and road conditions such as stairs or level ground, there is few method to calculate calorie consumption precisely by estimating these elements.

In this paper, we propose the method to estimate of daily life activities precisely and to calculate calorie consumption using the acceleration sensor on the smartphone.

2 RELATED WORK

2.1 State Estimation using Acceleration Sensor

There is a method of state estimation of daily activities using acceleration sensor [3]. This method makes multiple parameters from the data obtained from acceleration sensor, and estimates states from the characteristic values in each state. Figure 1 shows the variance of the acceleration, the power spectrum derived from FFT of the acceleration(FFT power spectrum) and the angle of the terminal in each state such as "Sitting", "Standing", "Walking" and "Running". These values are obtained from the acceleration sensor in the pocket of pants.

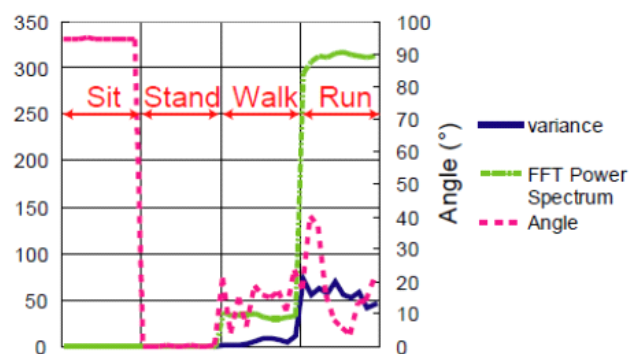


Figure 1: Variation of characteristic values

2.2 Measurement of Calorie Consumption

2.2.1 METs Values

There are many studies to apply the state estimation method to calculate calorie consumption using METs(Metabolic Equivalent of Tasks) values. The METs values [4] is the technique

proposed by the American College of Sports Medicine and is expressed in the following expression (1).

$$EE = 1.05 \times METs \times W \times T \quad (1)$$

EE means calorie consumption(kcal), W means the weight of the subject person(kg) and T means the duration of the activity(hours). The METs value means that the strength of activities is equivalent to several times at resting condition. The examples of METs values are the following table 1. For

Table 1: Examples of METs values

METs	Activiteis
1.0	Sitting quietly and watching television
1.2	Standing quietly
1.3	Sitting and reading books
1.4	Sitting and talking, eating
1.8	Standing and talking
3.0	Walking(67m/min, Level ground)
3.3	Walking(81m/min, Level ground)
3.8	Fast walking(94m/min, Level ground)
10.0	Running(161m/min, Level ground), Swimming
15.0	Running up the stairs

example, we can calculate calorie consumption on the following expression (2) when the subject person weighs 60kg and did normal walk of around 67m/min on level ground for 30 minutes.

$$1.05 \times 3.0 \times 60 \times 0.5 = 94.5(kcal) \quad (2)$$

2.2.2 Measurement of Calorie Consumption using State Estimation

The method considering the moving states [5] can estimate the moving state such as "Stopping" or "Walking" or "Running" or "Bicycle" or "Train, Car, Bus" using mobile terminals. Furthermore, in the "Walking" state this method distinguishes "Level ground", "Up the stairs" and "Down the stairs", and calculates calorie consumption from estimated state and METs. The state estimation of this method uses only the FFT power spectrum of the acceleration as the characteristic value. The measured power spectrums are converted into symbols decided beforehand. And each state is estimated from these plural symbols

The method considering daily life activities [6] can estimate the basic four states of daily life activities such as "Sitting" or "Standing" or "Walking" or "Running" using sensors on the market. The state estimation of this method uses the values of variance, dominant frequency, the FFT power spectrum of the acceleration and the angle of the terminal as the characteristic value.

2.2.3 Problems

Table 2 shows that the METs values are greatly different in the difference of speed and the road condition. This has a big influence on the calculated value of calorie consumption.

Table 2: Difference in METs between "Walking" and "Running"

METs	Activiteis
2.5	Walking(54m/min)
5.0	Walking(107m/min)
8.0	Walking down the stairs Running(134m/min)
10.0	Running(161m/min, Level ground)
15.0	Running up the stairs

Though the method of Ref. [5] considers only moving state, the estimation of "Bicycle" or "Train, Car, Bus" is unnecessary if we calculate calorie consumption only in daily life activities. This method does not consider moving speed. As for the method of Ref. [6], the road condition where the subject person "runs" or "walks" such as "up and down the stairs", is not estimated. For more accurate calculation of calorie consumption, it is necessary to be able to estimate these speeds and conditions.

3 PROPOSED METHOD

We propose the method to calculate calorie consumption in daily life activities. This method consists of two processes such as the process of state estimation and the process of calculation of calorie consumption. The process of state estimation uses the acceleration sensor of smartphone, and the process of calorie calculation uses METs values. And the basic daily life activities shall be classified into four state such as "Sitting", "Standing", "Walking" and "Running".

When the acceleration value is measured using smartphone, there is a problem that the acceleration values change by the wearing place of smartphone. In the proposed method, smartphone is attached to the pocket of pants(Fig.2). In this way, it



Figure 2: Wearing place of smartphone

becomes easy to distinguish the change of state by the angle of feet at the time of "Sitting" and "Standing". And at the time of "Walking" and "Running", it becomes easy to catch the change of the acceleration.

The frequencies of "Walking" and "Running" are usually delivered to 0-10Hz. Therefore the sampling rate of the acceleration sensor sets it to 20Hz. The state estimation process uses the variance of the acceleration, the basic frequency derived from FFT of the acceleration and the angle of the smartphone. The variance and the frequency of the acceleration are calculated from 64 latest acceleration data.

3.1 Process of State Estimation

The proposed method estimates the states of "Walking up the stairs", "Walking down the stairs" and "Running up the stairs" in addition to "Sitting", "Standing", "Walking" and "Running" states of the conventional method. And the method calculates the velocity of moving at the time of "Walking" and "Running". Figure 3 shows the overview of the proposed method.

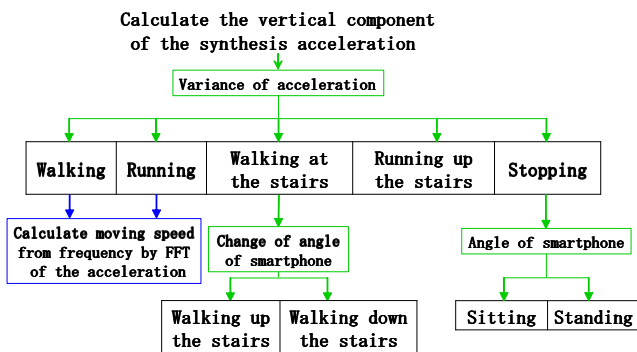


Figure 3: Overview of state estimation

Reference [3] shows that four states "Walking", "Running" and "Stopping" can be classified by the variance of the synthesis acceleration. Furthermore, the proposed method estimates the states "Walking on the stairs" or "Running at the stairs" in addition to the states "Walking" and "Running". However, we cannot distinguish the states "Walking on the stairs" and "Running at the stairs" from the states "Walking" and "Running" by the variance of the synthetic acceleration as shown in Fig. 4. The vertical movement of "Walking on

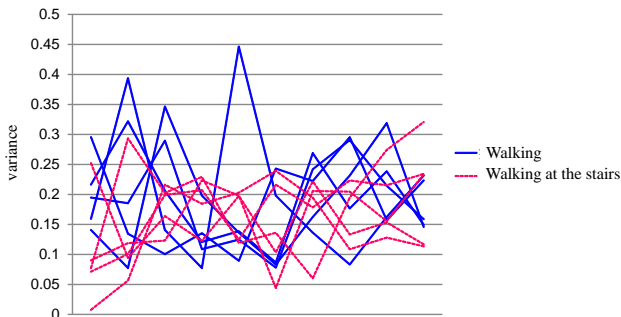


Figure 4: Variance of the synthesis acceleration

stairs" is stronger than normal "Walking". Therefore I verified whether we could detect the characteristic by the vertical component of the variance of the synthesis acceleration. Figure 5 shows the result. There are great difference of the

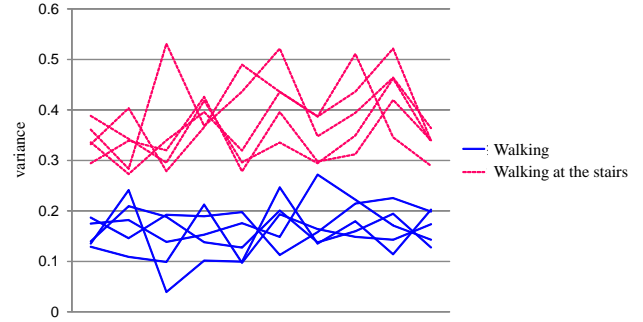


Figure 5: Variance of the vertical component of the synthesis acceleration

vertical component of variance of the acceleration, whether the subject person walks on the stairs, In this way we can estimate the states "Walking", "Running" "Walking on the stairs", "Running at the stairs" and "Stopping" only using the value of variance.

The calculation procedures of the vertical component of the synthesis acceleration are as follows.

1. About x-axis and y-axis and z-axis, the gravity component of the acceleration is removed from the acceleration of the x-axis and y-axis and z-axis.
2. The acceleration is composed from x, y, z component of acceleration that removed gravity component, according to the following expression.

$$Acc = \sqrt{x^2 + y^2 + z^2} \quad (3)$$

3. The angle θ is calculated from the inner product between the synthesis acceleration and the gravitational acceleration.

$$\theta = \cos^{-1}\left(\frac{\vec{a} \cdot \vec{b}}{|\vec{a}| |\vec{b}|}\right) \quad (4)$$

4. The vertical component of acceleration is calculated from the angle θ .

$$Acc_v = Acc \times \sin \theta \quad (5)$$

Acc means synthetic value of the acceleration. x , y and z mean x-axis(y-axis, z-axis) component of the acceleration that removed a gravity component respectively. θ means the angle between synthesis acceleration and the gravitational acceleration, \vec{a} means the acceleration that removed the gravity component, \vec{g} means the gravity component of the acceleration. Acc_v means the acceleration of the vertical direction.

3.1.1 State Estimation

Each state is estimated using each threshold found experimentally about the variance, the angle and the variation of the angle(differences between 64 latest maximums and minimum of the angle data) of measured acceleration. The setting of the threshold is described in section 3.1.3.

At first, basic state such as "Stopping", "Walking", "Walking on the stairs", "Running" and "Running up the stairs" are

estimated using the variance of the acceleration of the vertical direction. Table 3 shows the threshold. The state is decided where of the threshold range the variance va of measured acceleration is included in.

Table 3: Threshold of basic states

State	Threshold
Stopping	$v < 0.07$
Walking	$0.07 \leq v < 0.33$
Walking on the stairs	$0.33 \leq v < 0.53$
Running	$0.53 \leq v < 1.35$
Running up the stairs	$1.35 \leq v$

Then, when it is in the state of "Walking on the stairs", the states of "Walking up the stairs" and "Walking down the stairs" are estimated using the variation of the angle measured at smartphone. Table 4 shows the threshold. The state is decided where of the threshold range the variation of the angle a_v is included in.

Table 4: Threshold of "Walking on the stairs"

State	Threshold
Walking up the stairs	$0.67 < a_v$
Walking down the stairs	$a_v \leq 0.67$

Finally, when it is in the state of "Stopping", the states of "Sitting" and "Standing" are estimated using the angle data measured at smartphone. Table 5 shows the threshold. The state is decided where of the threshold range the angle data a is included in.

Table 5: Threshold of "Stopping"

State	Threshold
Sitting	$1.1 \leq a < 2.0$
Standing	$a < 1.1, 2.0 \leq a$

3.1.2 Calculation of the Velocity

The step is defined in Ref. [7], and calculates using the following expressions. S_w means stride in walking, S_r means the stride in running and H means the height of the subject person.

$$S_w(m) = H(m) \times 0.45 \quad (6)$$

$$S_r(m) = H(m) \times 0.5 \quad (7)$$

The velocity of the subject person v can be calculated on the following expression. S means the stride and P means the pace of walking/running.

$$v(m/min) = S(m/steps) \times P(steps/min) \quad (8)$$

The pace at the time of "Walking" or "Running" is calculated by FFT of the acceleration data.

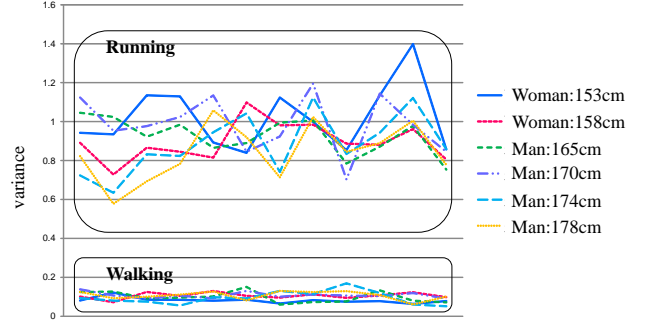


Figure 6: Variance of the acceleration at the time of "Walking" or "Running"

3.1.3 Setting of the Threshold

These thresholds are set using decision tree and the result of preliminary experiment. As a preliminary experiment, six subject people including man and woman continued each state for about 300 seconds. The threshold is set by the variance, the angle and the variation of the angle provided from this experiment using decision tree. Figure 6 shows the value of variance in each subject person at the time of "Walking" and "Running". Though the variance tends to become small if the subject person is short, the ranges of the variance are divided clearly at the time of "Walking" and "Running", and it is possible to distinct by the threshold.

3.2 Measurement of Calorie Consumption

The METs values corresponding to the state in the proposed method is set by table 6. At the time of "Walking" and

Table 6: METs values of each states

METs	Activities
1.5	Sitting
1.8	Standing
Expression(9)	Walking
Expression(10)	Running
3.0	Walking down the stairs
8.0	Walking up the stairs
15.0	Running up the stairs

"Running", METs values depend on the speed v and calculate on the following expressions by Ref. [3].

$$Walking : METs = 0.0272(m) \times v + 1.2 \quad (9)$$

$$Running : METs = 0.0930(m) \times v - 4.7 \quad (10)$$

4 EXPERIMENTAL EVALUATIONS

In this experiment, the proposed method is implemented on iPhone4. And the precision evaluation is divided into two evaluations. First one is the precision evaluation of the calculation of calorie consumption. Second one is the precision evaluation of the state estimation.

4.1 Precision Evaluation of the Measurement of Calorie Consumption

As an experiment scenario, six subjects move along the original route and measure the state of activities. The estimated precision is evaluated by comparison between the correct value and measured value by the proposed method and the measured value by the conventional method. The correct values are calculated from real activities and the METs values. The scenario is set as follows.

Table 7: The scenario of experiment

Order	Activiteis	Time
1	Walking	30sec
2	Walking up the stairs	30sec
3	Walking	30sec
4	Walking down the stairs	30sec
5	Standing	60sec
6	Running	15sec
7	Running up the stairs	30sec
8	Sitting	60sec

Figure 7 shows the result of the evaluation. As a result, the proposed method can calculate calorie consumption closer to the correct value than the conventional method.

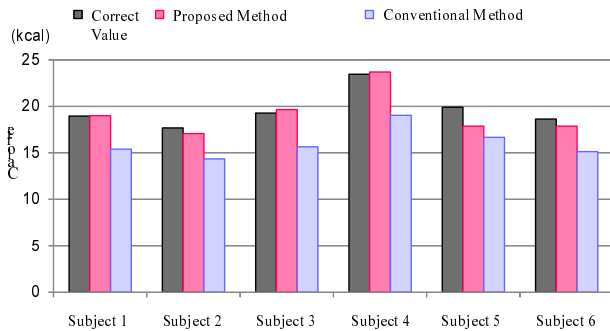


Figure 7: Result of evaluation of the calculation of calorie consumption

4.2 Precision Evaluation of the State Estimation

The correct answer rate of the state estimation process is calculated according to the following expressions and evaluates the precision of estimation. C means the correct answer rate, N_c means the number of states estimated correctly and N_e means the number of states estimated.

$$C(\%) = \frac{N_c}{N_e} \times 100 \quad (11)$$

The proposed method aims for highly precise calculation of calorie consumption at the time of the daily life activity by improving the precision of individual state estimation. Therefore it is desirable for the individual state estimation has around 90% of the correct answer rate.

Table 8 shows the result of the precision evaluation of the state estimation process. In this table, "W u s" means "Walking up the stairs", "W d s" means "Walking down the stairs" and "R u s" means "Running down the stairs". The vertical line expresses the real state and the horizontal line expresses the estimated result.

The precision of the state "Sitting" and "Standing" reached 100%, and "Walking" achieves 95.4%. On the other hand, The precision of the state "Walking up the stairs" achieves 70.7%, "Running" achieves 84.4% and "Running up the stairs" achieves 74.2%. The precision of the state "Walking down the stairs" turned out low about 44%.

4.3 Discussion

4.3.1 Discussion about Precision of the Measurement of Calorie Consumption

We consider about the proposed method's precision of the calculation of calorie consumption. From the result of Fig.7, the measured value of the our method is almost the correct value than the measured value of the conventional method. Therefore, our method can calculate calorie consumption precisely at the time of daily life activity than the conventional method. And the error average of the measured value of the proposed method and the correct value is about 0.69 kcal. On the other hand, the error average of the conventional method is about 3.6 kcal. These errors are errors per 4 minutes 30 seconds. If exchanging this for 18 hours that are the mean activity time of the person of the day, the error of the proposed method becomes equivalent to about 165 kcal. The error of the conventional method becomes equivalent to about 860 kcal. Therefore, the proposed method could reduce 86% of errors than the conventional method. 860 kcal which is the error of the conventional method is equivalent to 3-4 cups of rice bowls because one cup of rice bowl(140g) is 235 kcal [8]. 165 kcal which is the error of the proposed method does not reach one cup of rice bowl. However, this value is an error when exchanging the experiment scenario for 18 hours. Actually "Sitting" or "Standing" or "Walking" which can be estimated precisely by the proposed method is the major state of daily life activities. Therefore it is expected that the error becomes less than 165 kcal when the proposed method in daily life.

4.3.2 Discussion about Precision of the State Estimation

We consider about the proposed method's precision of the state estimation. From the result of table 8, "Sitting" and "Standing" and "Walking" can achieve more than 95% of high precision. In contrast, the estimated rate of "Walking down the stairs" is low with 44%. The "Walking down the stairs" causes many false estimates with the "Walking up the stairs" state. The reasons of increase of false estimations is the problem about the effectiveness of the thresholding. Figure 8 shows the change of the angle of smartphone at the time of "Walking down the stairs" and "Walking up the stairs". From the result, there are the parts which have a large difference of the change of the angle and the parts which have a small difference of the change of the angle. This depends on the individual difference in the way of going up stairs and going down stairs. In

Table 8: Result of evaluation of the state estimation

	Sitting	Standing	Walking	W u s	W d s	Running	R u s
Sitting	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Standing	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Walking	0.0%	0.0%	95.4%	0.0%	4.6%	0.0%	0.0%
W u s	0.0%	0.0%	19.5%	70.7%	0.0%	9.8%	0.0%
W d s	0.0%	0.0%	12.0%	32.0%	44.0%	12.0%	0.0%
Running	0.0%	0.0%	0.0%	0.0%	0.0%	84.4%	15.6%
R u s	0.0%	0.0%	0.0%	0.0%	3.2%	22.6%	74.2%

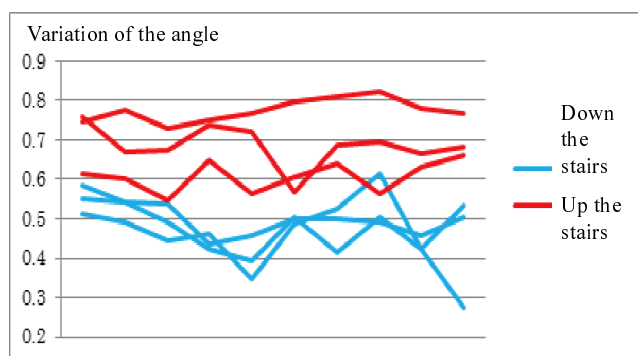


Figure 8: Variation of the angle

particular, at the time of "Walking up the stairs", the change of the angle grows larger at going up the stairs lifting feet up highly. Therefore, the correct answer rate of estimation can be improved by changing the threshold on individual basis dynamically. Figure 9 is one of the examples. Because the threshold is not suitable in the middle, the wrong state is estimated. By adjusting the threshold according to the personal characteristic, correct estimation is enabled as in the below graph of Fig. 9.

5 CONCLUSION

In this paper, we proposed the method to calculate calorie consumption precisely considering daily life activities. Our proposed method achieved precise calculation of calorie consumption by considering the road condition and movement speed.

We evaluated the precision of the proposed method by the comparison experiments between the conventional method and the proposed method. As a result, it was shown that the proposed method could calculate calorie consumption precisely in comparison with the conventional method. However, there is the room for improvement because the precision of the proposed method is low in "the case to walk down the stairs".

For future work, we need the improvement of the estimation accuracy of activities by applying plural learning models. In addition, it is necessary to evaluate the calculation of calorie consumption in the daily unit because the experiments of this paper depend on the original scenario.

REFERENCES

- [1] Japan Health Insurance Association, <http://www.kyoukaikenpo.or.jp/>.
- [2] J. A. Levine, L. M. Lanningham-Foster, S. K. McCrady, A. C. Krizan, L. R. Olson, P. H. Kane, M. D. Jensen, and M. M. Clark, Interindividual variation in posture allocation: possible role in human obesity, *Science*, Vol. 37, No. 5709, pp. 584–586 (2005).
- [3] H. Kurata, Y. Kawahara, H. Morikawa, and T. Aoyama, User Posture and Movement Estimation Based on 3-Axis Acceleration Sensor Position on the User's Body, *IPSJ SIG Technical Reports*, Vol. 2006, No. 54(2006-UBI-011), pp. 15–22 (2006).(*in Japanese*)
- [4] Ministry of Health, Labour and Welfare, Exercise Guide 2006 (2006).(*in Japanese*)
- [5] A. Minamikawa, A. Kobayashi, and H. Yokoyama, Energy Expenditure Monitoring System on Mobile Phone Using Information Gain Based Locomotion Estimation Method, *IPSJ Journal*, Vol. 52, No. 2, pp. 866–876 (2011).(*in Japanese*)
- [6] N. Ryu, Y. Kawahara, A. Kobayashi, and T. Asami, A Energy Expenditure Estimation Method for Non-Exercise Activity Thermogenesis Using Accelerometer, *IPSJ SIG Technical Reports*, Vol. 2008, No. 40(2008-UBI-018), pp. 67–74 (2008).(*in Japanese*)
- [7] Run & Walk, <http://run.auone.jp/>.
- [8] Calories in Japanese foods, <http://www.eiyoukeisan.com/JapaneseFoodCalorie/>.
- [9] A. Kobayashi, T. Iwamoto, and S. Nishiyama, Shaka : Method for Estimating User Movement Using Mobile Phone, *IPSJ SIG Technical Reports*, Vol. 2008, No. 44(2008-MBL-045), pp. 115–120 (2008).(*in Japanese*)
- [10] J. Sekiguchi, T. Matsui, and Y. Niitsu, Seated Position Context Presumption Method, *Proceedings of the 2010 IEICE Tokyo Branch Students' Conference*, p. 108 (2010).(*in Japanese*)
- [11] Y. Matsumura, K. Hirobe, K. Nishino, Y. Yamanaka, and T. Nakamura, Physical activity measurements based on 3-axis acceleration method, *Matsushita Electric Works technical report*, Vol. 56, No. 2, pp. 67–72 (2008).(*in Japanese*)
- [12] Aoyama Clinic, http://www1.s3.starcat.ne.jp/aoyama_c/.
- [13] Weka, <http://www.cs.waikato.ac.nz/ml/weka/>.
- [14] M. St-Onge, D. Mignault, D. B. Allison, and R. Rabasa-Lhoret, Evaluation of a portable device to measure daily

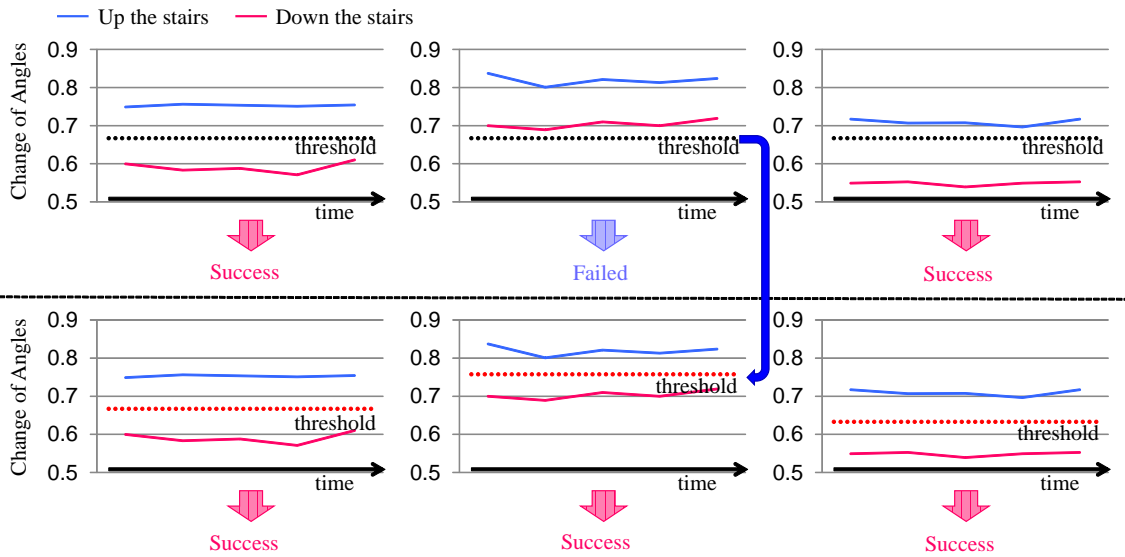


Figure 9: Change of the dynamic threshold

energy expenditure in free-living adults, The American Journal of Clinical Nutrition, Vol. 85, No. 3, pp. 742–749 (2007).

User's Communication Behavior in the Pseudo Same-room Videoconferencing System

Mamoun Nawahdah* and Tomoo Inoue**

*Faculty of IT, Computer Science Department, Birzeit University, Birzeit, Palestine

**Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Japan
 *nawahdah@birzeit.edu, **inoue@slis.tsukuba.ac.jp

Abstract - This paper presents user's communication behavior in the pseudo same-room videoconferencing named "Being Here System," in comparison with a conventional videoconferencing. The system extracts the remote person's figure and superimposes it on the local site's front view in a large display. In this way, the local person may feel as if the remote person was before him/her in his/her spatial environment. To investigate the influence of the system on user's communication, the recorded video of the system evaluation experiment was analyzed. This revealed that the system significantly affected user's communication behavior such as turn taking, speech overlapping, total speech time, and gaze directions.

Keywords: Videoconferencing; presence; video overlay; telecommunication; Kinect application.

1 INTRODUCTION

Communication is a very important link with others, if we communicate well with others, we are able to better understand what others around us want, need, expect of us, and what they are able to do and likewise, they will understand what we want, need, etc. Communication can be carried out in face-to-face (F2F) or through media. In F2F communication, the exchange of information, thoughts, and feelings is made when the participants exist in the same physical space at the same time. In this communication, nonverbal cues (e.g., eye contacts, facial expressions, appearances, body movements, interpersonal distances, etc.) may influence the way the message is interpreted by the receiver. In contrast, although mediated communication including videoconferencing provides people with many advantages given the increased globalization and the need for rapid knowledge transfer across borders and time-zones, the absence of nonverbal cues and tacit knowledge transfer may make communication difficult. Hence communication process is affected in mediated communication. A person may feel less presence of remote participants in mediated settings, and he/she may fail to correctly and/or accurately interpret other people's behavior. Therefore, one of the design goals of a videoconferencing system is to create a medium setup that is as close as possible to F2F.

Many studies have suggested that generating a life-sized view is likely to enhance the user's sense of presence [8, 12, 13, 24, 11, 16]. Here "presence" or "sense of presence" refers to the user's feeling of connection to the remote person with whom they are interacting [18]. The life-sized view makes it easy to read the other person's behavior such

as eye movements, facial expressions, gestures, and postures, which are essential for smooth communication.

Large displays can be used to achieve a life-sized view. However, typically, this will mean that a considerable region of the local person's front view will be replaced by the remote site's background, and this may decrease the user's sense of presence, as there is no integration or continuity in the local person's front view. Moreover, the remote site's background in some environments might be 'cluttered' with static or movable objects. This may either be a distraction or be more engaging, giving a greater sense of the other person's environment [5].

"Being Here System (BHS)" is a system to achieve pseudo same-room videoconferencing system using a large display [23]. The system provides the communication environment, in which the remote user's life-sized figure is visually situated in the local site (Figure 1) and vice versa. The display shows local site's front view, which would otherwise have been obstructed by the display, as a background. In this way, user feels as if remote user is present before him/her in the same room. In other words, the user feels co-presence of the other remote user. BHS was evaluated by a questionnaire filled by users after performing a videoconferencing experiment. The questionnaire results revealed that BHS achieved higher sense of co-presence of remote users, in comparison with a conventional videoconferencing system.



Figure 1: A user talks to a remote user through BHS.

In this paper, we further investigated the user's behaviors when communicating using BHS. The motivated question is whether BHS affects verbal and/or nonverbal communication structure. The considered verbal communication parameters in this study are: turn taking, speech time, and speech overlapping. Regarding nonverbal parameters, person's gaze direction is considered. The user's

communication behaviour analysis revealed that BHS has significantly affected users' conversational behaviours as if participants were in the same room¹.

2 RELATED WORKS

2.1 Media Space Systems

There have been various studies done on remote communication and media spaces, and a host of systems have been developed over time. Many of these studies have been devoted to proposing and/or implementing methods aimed at enhancing the sense of presence in videoconferencing.

One early system called "Hydra" sought to enhance the sense of presence by supporting directional gaze cues and selective listening in 4-way videoconferencing [27]. A multi-party videoconferencing system called "MAJIC" was constructed by Okada et al. to support eye contact [24]. In this system, life-sized video images of participants were projected onto a large curved transparent display. Another line of research focused on the seating arrangement in video-mediated meetings, in order to enhance the sense of presence [12]. The system was designed for multiple participants so that the video image of any remote participant be always placed where a viewer need to make no effort to see it. A different approach to enhance the sense of presence was introduced by Morikawa et al. [20]. In this study, a system called "HyperMirror" was constructed, in which all participants were meant to feel as if they were sharing the same virtual space. To provide a greater sense of presence than had been achieved with conventional desktop videoconferencing, Gibbs et al. created the "TELEPORT" system [8], which was based on special rooms, called display rooms, in which one wall was a "view port" into a virtual extension. A side-by-side media space concept was proposed to enhance the presence feelings, which was suggested to be more appropriate for side-by-side style interactions such as collaborative writing and training [28]. Other effective attempts to enhance the presence feelings involved movable displays [22] and movable cameras [21].

It is natural to devote more attention to people present before one, since the felt presence of remote people is considerably weaker [32]. To overcome this inclination, robotic means have been employed to convey the sense of presence in videoconferencing, enhancing the remote people's felt presence. In this regard, a study by Sakamoto et al. investigated the effect of using a humanoid robot system as a telecommunication medium [26]. Another study, by Yankelovich et al., introduced a system called "Porta-Person" to enhance the sense of social presence for remote-meeting participants [32]. This goal was achieved by providing a high-fidelity audio connection and a remotely controlled telepresence display with video or animation. In the same manner, Venolia et al. developed a telepresence device, called "Embodied Social Proxy (ESP)", which represented a remote coworker at roughly human-scale [29]. In this system, they found that the physical presence of the

ESP was a powerful reminder of the presence of the remote worker in the meetings.

The studies and implemented systems above focused primarily on creating a high-presence media space. Our study, in turn, makes its own contribution to this field. To mimic real situations, the remote person's figure should be presented locally, without his/her remote background. Typically, this can be achieved by using mixed-reality (MR) technology and special head-mounted display (HMD) equipment [11, 4, 15]. Using HMD for some people might be encumbering and uncomfortable. This setup is likely to decrease the sense of presence. In contrast, our proposed system can be easily implemented in both sites, allowing both participants to experience the same effects.

2.2 Commercial Videoconferencing

In commercial videoconferencing business firms, many solutions have been introduced under the name "Telepresence" technology for high presence feelings. Telepresence is defined as an illusion that a mediated experience is not mediated [18]. In videoconferencing experience, telepresence gives you the feeling as if the remote participants are in the same room with you. To create the same-room illusion, some commercial telepresence solutions use a combination of technology elements, such as utilizing large displays for life-sized dimensions and hidden high-definition cameras strategically placed to create the appearance of a direct eye contact, and environmental design, such as consistent furniture arrangements across locations. The life-sized dimensions allow participants to see facial expressions, make eye contact, and read body language. Such solutions are: Cisco TelePresence TX9000 Series², Polycom® RealPresence™ Immersive³, TANDBERG⁴, PeopleLink TelePresence⁵, etc.

In one hand these solutions simulate high presence meeting environments as if the other people were sitting across the table in the same room. But on the other hand these solutions are very expensive, require large-spaces, and have to be installed in a fixed environment with pre-installed matching furniture in both sides to achieve maximum telepresence feelings. In contrast, BHS can be implemented using an affordable equipments and can be installed easily almost anywhere.

2.3 Verbal and Nonverbal Communication Analysis

It's well known that in F2F communication, people switch speaking and listening by using a complicated mechanism of verbal and nonverbal cues [2]. A major nonverbal cue in speaking involves the use of eye contact [1]. In F2F communications, failure to maintain eye contact is commonly considered to be a sign of deception, and leads to feelings of mistrust [2]. Vertegaal et al. concluded that

¹ This research was partially supported by the JSPS Grant-in Aid for scientific research 22500104 and 23500158.

² <http://www.cisco.com/en/US/products/ps12453/>

³ http://www.polycom.com/products/telepresence/_video/

⁴ <http://www.tandberg.com/>

⁵ <http://www.peoplelink.in/telepresence.html>

gaze is an excellent predictor of conversational attention in multiparty conversations [30]. A study by Karmer et al. proposed a method of measuring people's sense of presence in videoconferencing system based on linguistic features of their dialogues [17]. This study shows that 30% of the variance in self-reported presence can be accounted for by a small number of task-independent linguistic features.

The seating arrangements on group video communication affect participant's behaviors as well. A study by Inoue et al. presented a videoconferencing system "HERMES" that integrates F2F and video-mediated meetings [12]. In This study they observed that participants tended to pay much attention to the monitor when using lined-up seating arrangement. This problematic behavior solved by the combination of round seat arrangement and multiple monitors. Another study by Yamashita et al. revealed that seating arrangements affect speaker switches without verbal indication of the next speaker [31]. This study found that in some seating arrangement, the participants shared a higher sense of unity and reached a slightly better group solution.

Our study as well examined the proposed high-presence videoconferencing system for any verbal and/or nonverbal effects on communication comparing with a conventional videoconferencing system.

3 BEING HERE SYSTEM

A videoconferencing system "Being Here System (BHS)" was constructed to achieve pseudo same-room environment to the users. The current system consisted of two isolated sites, 'Site A' and 'Site B'. The two sites were connected over a local network to permit the exchange of live video. Each site was equipped with a display installed upright 70 cm above the floor, a USB camera fixed behind the display, a Kinect™ depth camera, a computer connected to the network, two speakers and a microphone, and a chair. The user was seated at 1.2 m distance from the display since this was considered to be appropriate distance for F2F meetings [10].

The processing of BHS is shown in Figure 2. We used a USB camera to capture the local site's front view, that is, the region concealed behind the display. The USB camera was placed behind the display in the center, and the camera's angle and zoom were calibrated so that the area behind the display was exclusively captured. This captured image (640 by 480 pixels) was used as a background for the display.

To capture the site view and extract the user's figure from it at run-time, we used a Kinect depth camera. The Kinect was placed centrally over the display and focused on the person's face. This is considered to be a best placement given the constraints of the environment [5]. OpenNI API is used to analyze the Kinect image depth data by identifying the user in the scene and replacing the background with a transparent color. The resulted image is transmitted to the other site at 15 frames per second rate.

The final step in the process was to superimpose the received remote user figure onto the local front view. This was accomplished by merging the extracted user's figure and the background. Finally, the resulting view (640 by 480 pixels) was presented on a large display.

With this simple system architecture, the system can be easily expanded to the multi-point distributed conferencing system that connects three or more sites. This is a noteworthy feature that other existing systems have not achieved because of their limited spatial alignment and/or expensive customized devices.

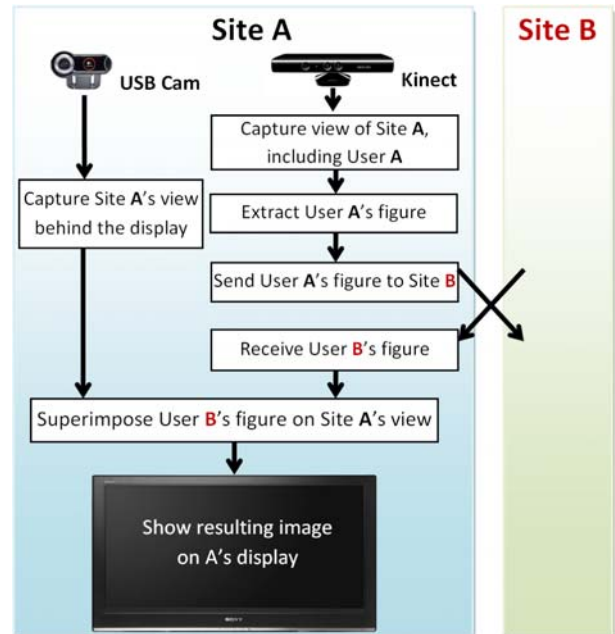


Figure 2: The process diagram of BHS.

4 SYSTEM EVALUATION

An experiment was conducted to evaluate BHS. The main objective of this experiment was to study the influence of using BHS on user's communication behaviour, in comparison with conventional videoconferencing. Two remote sites were constructed (Site A and B). In Site A, a large flat-panel display (46 inches) was used (Figure 1), while in Site B, a 30-inch display was used. The 30-inch display was fixed in a vertical portrait position, presenting a life-sized image of an adult's upper body (Figure 3). We used the portrait mode to study the effects of different background sizes if any.



Figure 3: The 30-inch portrait display setup.

4.1 Conditions

Two videoconferencing modes were established to enable the participants to communicate with each other: “*Conventional mode*”, in which the remote site view was displayed as is in the local site's display; and “*Superimpose mode*”, in which the remote person's figure was extracted and superimposed on the local site's front view.

In this experiment, we considered the following videoconferencing conditions, to evaluate the constructed system:

- Large Superimpose (*LS*): superimpose mode via large display.
- Portrait Superimpose (*PS*): superimpose mode via portrait display.
- Large Conventional (*LC*): conventional mode via large display.
- Portrait Conventional (*PC*): conventional mode via portrait display.

4.2 Participants

A total of 18 participants took part in the experiment, 7 females and 11 males. The participants' ages ranged from 23 to 36 years old, most were undergraduate or graduate students from the same university, and 17 participants had had previous experience using videoconferencing systems. Most used the videoconferencing principally to talk to a remote family member and/or remote close friend. The participants were divided into groups of two. Within each group, we made sure that the participants were familiar with each other as we wanted the interaction to be as smooth as possible.

4.3 Procedure

In each experiment, a group of two participants were recruited to perform videoconferencing tasks with each other. One of the participants used the system at Site A, while the other used Site B. Before performing the videoconferencing tasks, the participants were asked to complete a basic demographic survey. After this, the researcher introduced the system to the participants. The experiment began with a familiarization session for five minutes. Each participant performed four videoconferencing sessions to test the conditions (two sessions at Site A, and two sessions at Site B). In each session, participants were instructed to talk about a selected general topic for an average of 10 minutes. After that they were asked to complete the questionnaire about the system they experienced, independently of each other. The four general topics were:

- Study life in X city: discuss with the other person the pros and cons of studying in X city; how long you have been in X city; why you choose X university, compare X city with other cities you have been in, etc.
- Buying a new laptop: discuss the laptop's specifications; the suggested shops; prices; usage; etc.
- Planning a trip: for the coming summer vacation, discuss the trip's options; where to go; locally or abroad; cost; weather; attraction; etc.

- What your plans after graduation: discuss with the other person your plans after graduation, the possibility of pursuing a higher degree; work options, etc.

The conditions orders were counterbalanced across participants, to ensure that the order of the tested conditions would not affect the result.

4.4 Questionnaire

In the questionnaire, we asked participants to evaluate each of the statements according to the feeling they experienced during the videoconferencing session. The principal aim of the study was to assess participants' sense of the other person's co-presence while using the constructed media space. To investigate the participants' sense of co-presence in each condition, the following statements were used: [10]

- *Q1*: “I felt as if the other person existed in the same room.”
- *Q2*: “I didn't feel as if I were talking with the other person in the same room.”
- *Q3*: “I felt as if I were facing the other person in the same room.”

The feeling of spatial distance between the participants is one aspect of the sense of co-presence. In this research, our focus was limited to the social distance communication range, which is best for business meeting activities and social interactions. We assumed that participants who experienced the superimpose mode would feel as if the other person were much closer, almost as if 'here', in comparison with the conventional mode. To evaluate this aspect, the following statement was used:

- *Q4*: “I felt that the distance between me and the other person was comfortable for chatting.”

In addition, we asked the participants to roughly estimate the distance between themselves and the other person while videoconferencing:

- *Q5*: “I felt that the distance between me and the other person was around: _____ “

All of these statements, except *Q5*, were rated on a 9-point Likert scale for precise assessment, where 1 = strongly disagree, 3 = disagree, 5 = neutral, 7 = agree, and 9 = strongly agree.

4.5 Session Recording Setup

Two cameras were used to record the experiment sessions at HD 720 resolution (1280 by 720 pixels). The first camera was placed over the display facing the participant in order to capture his/her facial expressions, gestures, and postures. The second camera was installed upright 1 m above the floor beside participant in order to capture him/her from the left side and the display content.

5 RESULTS

5.1 Questionnaire Data

Figure 4 shows the average results of the participants' sense of the other person's presence while videoconferencing, under the four conditions. A comparison was done using a two-factor ANOVA test. The first factor is the videoconferencing mode (i.e. Conventional and Superimpose). The second factor is the used display (i.e. Large and Portrait). We found a main effect of videoconferencing mode over the participants' sense of other person's presence as if in the same room, (Q1: $F(1,68) = 55.26$, $p < 0.01$), (Q2': $F(1,68) = 14.08$, $p < 0.05$), and (Q3: $F(1,68) = 31.71$, $p < 0.01$). This indicates that the superimposed videoconferencing mode enhanced the presence feelings more than the conventional videoconferencing mode. On the other hand, the results shows no main effect of the used display over the participants' sense of other person's presence as if in the same room, (Q1: $F(1,68) = 0.31$), (Q2': $F(1,68) = 0.0$), and (Q3: $F(1,68) = 0.06$). The result also shows that there is no interaction between the used mode and display over the presence feelings, (Q1: $F(1,68) = 2.31$), (Q2': $F(1,68) = 2.73$), and (Q3: $F(1,68) = 1.34$).

Moreover, we found a main effect of videoconferencing mode over the feeling of comfortable distance between the user and the other person (Q4: $F(1,68) = 7.14$, $p < 0.01$), while no main effect of the used display was reported (Q4: $F(1,68) = 0.89$). The result also shows that there is no interaction between the used mode and display over the distance, (Q4: $F(1,68) = 0.62$). This indicates that the superimposed videoconferencing mode enhanced the feeling of comfortable distance between the user and the other person. In addition, participants who used the superimpose videoconferencing mode were able to estimate the distance more accurately. The average estimated distance was as follows:

- LS: 1.3 m (s.d. = 0.6).
- PS: 1.3 m (s.d. = 0.6).
- LC: 2.2 m (s.d. = 1.2).
- PC: 1.8 m (s.d. = 1.0).

(The actual distance between the participant and the display was 1.2 m).

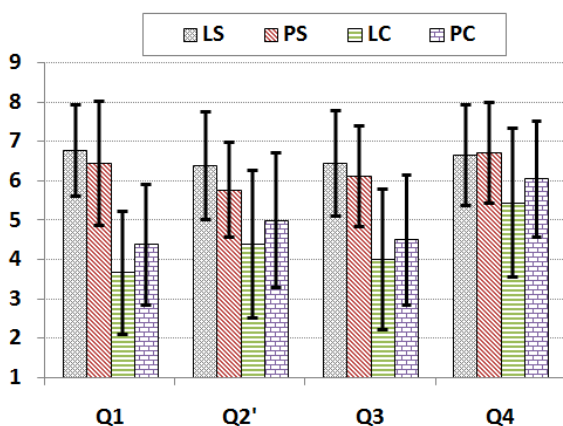


Figure 4: The participants' average sense of presence results. (Note: Q2' result is the positive form of the original Q2).

5.2 Analysis of Video Data

Communication behavior was analyzed using video data. The analysis was made from quantitative viewpoint. We are aware qualitative viewpoint such as what were talked and how they were talked could add more results, but that is future work.

ELAN¹ tool was used to annotate the recorded video of experiment sessions. A total of 36 recorded video (9 groups by 4 conditions) were annotated for user's communication behaviors such as talking and gaze. Only the middle 2 minutes of each session were analyzed at this stage of analysis (a total of 72 minutes of data). Figure 5 shows a screenshot of one of the ELAN's annotated video.

The following terms were used in annotating and analyzing the recorded video sessions:

- **Talking**: happens when a person speaks for at least 1.5 seconds [14].
- **Turn Taking**: In conversation analysis, turn taking term is defined as the manner in which orderly conversation normally takes place. The principles of turn-taking were first described by sociologists Sacks et al. in [25]. In this study, we adopted the same turn definition from [27] as the person's number of continuous segment of speech between silent intervals for at least 1.5 seconds.
- **Overlapping**: is a simultaneous speech by two persons. This might happen as talking turn change or as simultaneous reply to other person's speech while talking.
- **Gaze**: happens during a conversation when two people look at one another [1].
- **Gaze-off**: this term is defined for the analysis in this paper. It happens when the person avert his/her gaze from other person.



Figure 5: Screenshot of one of the ELAN's annotated video.

5.3 Talk Analysis

Figure 6 shows the average results of the participants' number of turn taking while videoconferencing, under the four videoconferencing modes. A comparison was done under the four conditions using a one-way repeated-measures ANOVA test. We found a significant difference in number of turn taking ($F(3,51) = 6.49$, $p < 0.05$). A Tukey's

¹ <http://tla.mpi.nl/tools/tla-tools/elan/>

HSD post-hoc test was performed in order to determine which condition's mean was different from the others. For this aspect we found that the superimpose conditions were significantly different from the conventional conditions.

Figure 7 shows the average results of participants' percentage of individual speech while videoconferencing. For this aspect, we found no significant difference between the tested conditions ($F(3,51) = 0.48$).

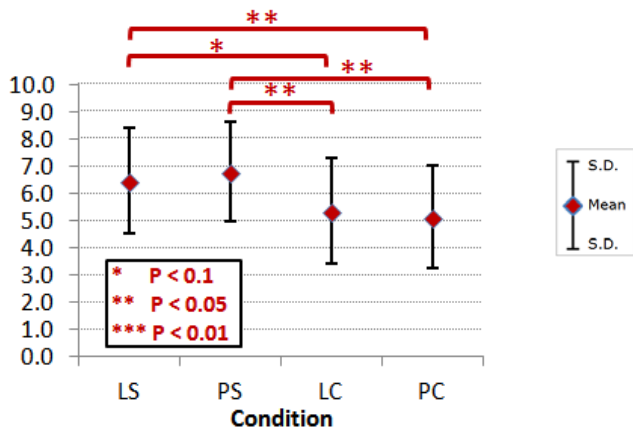


Figure 6: The participants' average number of turn taking per minute.

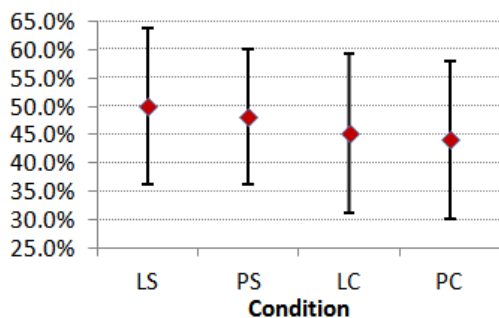


Figure 7: The participants' average percentage of speech per minute.

Figure 8 shows the average results of the participants' percentage of overlapping talk while videoconferencing. We found a significant difference in percentage of overlapping ($F(3,51) = 11.69, p < 0.01$). For this aspect we found that both the superimpose conditions were significantly different from the conventional conditions.

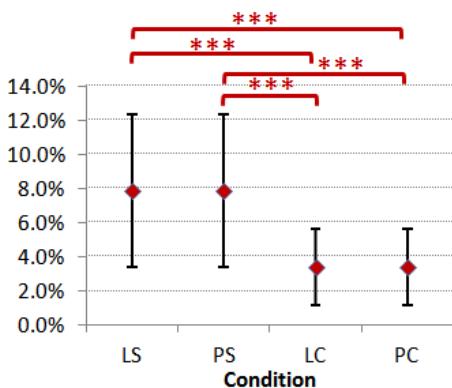


Figure 8: The participants' average percentage of talk overlapping per minute.

5.4 Gaze Analysis

Figure 9 shows the average results of the participants' number of gaze-off while videoconferencing. We found a significant difference in number of gaze-off ($F(3,51) = 4.83, p < 0.05$). For this aspect we found that both the superimpose conditions were significantly different from the conventional conditions.

6 DISCUSSION

We expected that the results could be different depending on the display sizes because of the different background sizes. One participant mentioned that the portrait display's wide border consumed a considerable amount of the front view compared with the large display, which may be related to the study by Bi et al. on the effects of bezels of large tiled display that the bezels affected tunnel steering [3]. But the results were almost very close over the same videoconferencing mode. In the large superimpose condition; a large part of the front view was displayed as a background. While in the portrait superimpose condition; a small part of the front view was displayed as a background. However, in this condition the participants were able to see more actual front view around display. This indicates that the display size has no major effect as long as the displayed background is integrated with the actual front view.

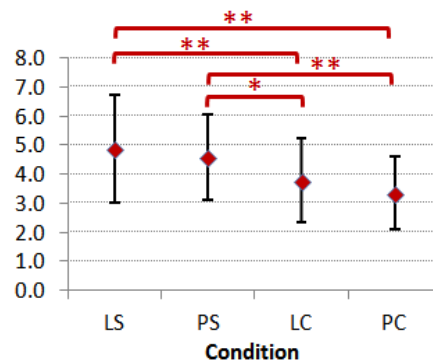


Figure 9: The participants' average number of gaze-off per minute.

The superimpose videoconferencing mode, using either large or portrait display, improved participants' sense of co-presence of the other person, in comparison with the conventional mode. This result supports our assumption that displaying the remote site's life-sized figure alone (after removing the remote site background) would increase the realism of the view and the sense of the other person's co-presence.

The superimpose mode simulated an actual F2F communication configuration. Some participants noted that the superimposition of the other person's figure onto the local front view made it appear as if they were seeing the other person in three dimensions. Some of the participants enjoyed the superimposing mode, with some attempting to

shake hands with each other through the system, and others moving and rotating and asking the other person to do likewise. In the conventional mode, nothing curious in participants' behavior was noticed.

Regarding the estimation of the distance between the participants during the videoconferencing sessions, the superimpose mode tended to result in a better distance assessment, compared to the conventional mode, possibly owing to the fact that participants managed to compare the relative distance of the remote person's figure with respect to the local site's front view shown in the display.

Furthermore, in this paper we show that the superimpose mode significantly affected participants' verbal communication. The results revealed that under superimpose conditions more speaking turns were taken compared with conventional conditions (Figure 6). The superimpose mode increased the number of turns by 1.4 more than the conventional mode. The participants' average percentage of speech wasn't affected by the tested conditions (Figure 7). The result shows that each participant talked for 48% of the session time on average. This result is consistent with a related research by Sellen [27]. We found that the percentage of speech overlapping in the superimpose conditions were twice more than the conventional conditions (Figure 8). Because Cohen's study [7] and Sellen's study [27] found that face-to-face imposes more simultaneous talk compared with video conditions, we can conclude that our proposed superimpose mode was closer to the F2F in this aspect.

Gaze awareness has been shown to be an important aspect of nonverbal communication [6, 9, 27]. In our study, we investigated the gaze-off (avert) aspect while communicating under the tested conditions. The result shows that the participants tended to avert their gazes more when they used the superimposed conditions compared to the conventional conditions (Figure 9). In F2F conversations, people use more gaze when they are further apart [1]. This means that the participants who used the superimposed conditions might feel closer to the other person.

7 CONCLUSION AND FUTURE WORK

In BHS, the pseudo same-room effect is achieved by superimposing the remote person's figure, which is extracted from the remote site view using a Kinect depth camera, with the local front view on a large display. BHS effectively reduced the psychological distance between the remote participants.

In this study, we investigated user's verbal and nonverbal communication behaviors while using BHS, in comparison with a conventional videoconferencing. The analysis of the recorded video of the system use revealed that using BHS significantly affected user's communication behavior. This result suggests that considering the local site front view as a background is one practical way to create the same-room illusion, which facilitates communication.

REFERENCES

- [1] M. Argyle, *Bodily Communication*, 2ed. Routledge, (1988).
- [2] E. Bekkering, and J. Shim, Trust in videoconferencing. *Commun. ACM*, Vol. 49, No. 7, pp.103–107 (2006).
- [3] X. Bi, S. Bae, and R. Balakrishnan, Effects of interior bezels of tiled-monitor large displays on visual search, tunnel steering, and target selection, In *Proceedings of the 28th international conference on Human factors in computing systems, CHI'10*, ACM, pp.65–74 (2010).
- [4] M. Billinghamurst, and H. Kato, Out and about real world teleconferencing, *BT Technology Journal*, Vol. 18, pp.80–82 (2000).
- [5] D. Chatting, J. Galpin, and J. Donath, Presence and portrayal: video for casual home dialogues, In *Proceedings of the 14th annual ACM international conference on Multimedia, MULTIMEDIA'06*, ACM, pp.395–401 (2006).
- [6] M. Chen, Leveraging the asymmetric sensitivity of eye contact for videoconference, In *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves, CHI'02*, ACM, pp.49–56 (2002).
- [7] K. Cohen, Speaker interaction: video teleconferences versus face-to-face meetings, In *Proceedings of Teleconferencing and Electronic Communications*, pp.189–199 (1982).
- [8] S. Gibbs, C. Arapis, and C. Breiteneder, Teleport towards immersive copresence, *Multimedia Syst.* 7, pp.214–221 (1999).
- [9] D. Grayson, and A. Monk, Are you looking at me? eye contact and desktop video conferencing, *ACM Trans. Comput.-Hum. Interact.* Vol. 10, No. 3, pp.221–243 (2003).
- [10] Y. Ichikawa, K. Okada, G. Jeong, S. Tanaka, and Y. Matsushita, Majic videoconferencing system: experiments, evaluation and improvement, In *Proceedings of the fourth conference on European Conference on Computer-Supported Cooperative Work, ECSCW'95*, Kluwer Academic Publishers, pp.279–292 (1995).
- [11] T. Inoue, Mixed reality meeting system enabling user to keep and share interpersonal distance in the real world, *Journal of Information Processing Society of Japan*, Vol. 50, No. 1, pp.246–253 (2009).
- [12] T. Inoue, K. Okada, and Y. Matsushita, Integration of face-to-face and video-mediated meetings: Hermes, In *Proceedings of the international ACM SIGGROUP conference on Supporting group work: the integration challenge, GROUP'97*, ACM, pp.405–414 (1997).
- [13] T. Ishida, A. Sakuraba, and Y. Shibata, Proposal of high realistic sensation system using the large scale tiled display environment, In *Proceedings of Network-Based Information Systems (NBIS)*, pp.444–449 (2011).
- [14] J. Jaffe, and S. Feldstein, *Rhythms of Dialogue*, Academic Press, New York, NY, USA, (1970).

- [15] T. Kantonen, C. Woodward, and N. Katz, Mixed reality in virtual world teleconferencing, In 2010 IEEE Virtual Reality Conference (VR), pp.179–182 (2010).
- [16] E. Koh, Conferencing room for telepresence with remote participants, In Proceedings of the 16th ACM international conference on Supporting group work, GROUP'10, ACM, pp.309–310 (2010).
- [17] A. Kramer, L. Oh, and S. Fussell, Using linguistic features to measure presence in computer-mediated communication, In Proceedings of the SIGCHI conference on Human Factors in computing systems, CHI'06, ACM, pp.913-916 (2006).
- [18] M. Lombard, and T. Ditton, At the heart of it all: The concept of presence. *Computer-Mediated Communication*, Vol. 3, No. 2, (1997).
- [19] M. Mantei, R. Baecker, A. Sellen, W. Buxton, T. Milligan, and B. Wellman, Experiences in the use of a media space, In Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology, CHI'91, ACM, pp.203–208 (1991).
- [20] O. Morikawa, and T. Maesako, Hypermirror: toward pleasant-to-use video mediated communication system, In Proceedings of the 1998 ACM conference on Computer supported cooperative work, CSCW'98, ACM, pp.149–158 (1998).
- [21] H. Nakanishi, K. Kato, and H. Ishiguro, Zoom cameras and movable displays enhance social telepresence, In Proceedings of the 2011 annual conference on Human factors in computing systems, CHI'11, ACM, pp.63–72 (2011).
- [22] H. Nakanishi, Y. Murakami, and K. Kato, Movable cameras enhance social telepresence in media spaces, In Proceedings of the 27th international conference on Human factors in computing systems, CHI'09, ACM, pp.433–442 (2009).
- [23] M. Nawahdah, and T. Inoue, Being Here: Enhancing the Presence of a Remote Person through Real-Time Display Integration of the Remote Figure and the Local Background, *Transactions of the Virtual Reality Society of Japan*, Vol. 17, No. 2, pp.101-109 (2012).
- [24] K. Okada, F. Maeda, Y. Ichikawa, and Y. Matsushita, Multiparty videoconferencing at virtual social distance: Majic design, In Proceedings of the 1994 ACM conference on Computer supported cooperative work, CSCW'94, ACM, pp.385–393 (1994).
- [25] H. Sacks, E. A. Schegloff, and G. A. Jefferson, Simplest Systematics for the Organization of Turn-Taking for Conversation, *Language* Vol. 50, No. 4, pp. 696–735 (1974).
- [26] D. Sakamoto, T. Kanda, T. Ono, H. Ishiguro, and N. Hagita, Android as a telecommunication medium with a human-like presence, In Proceedings of the ACM/IEEE international conference on Human-robot interaction, HRI'07, ACM, pp.193–200 (2007).
- [27] A. J. Sellen, Speech patterns in video-mediated conversations, In Proceedings of the SIGCHI conference on Human factors in computing systems, CHI'92, ACM, pp.49–59 (1992).
- [28] P. Tanner, and V. Shah, Improving remote collaboration through side-by-side telepresence, In Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems, CHI EA'10, ACM, pp.3493–3498 (2010).
- [29] G. Venolia, J. Tang, R. Cervantes, S. Bly, G. Robertson, B. Lee, and K. Inkpen, Embodied social proxy: mediating interpersonal connection in hub-and-satellite teams, In Proceedings of the 28th international conference on Human factors in computing systems, CHI'10, pp.1049–1058 (2010).
- [30] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt, Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes, In Proceedings of the SIGCHI conference on Human factors in computing systems, CHI'01, ACM, pp.301–308 (2001).
- [31] N. Yamashita, K. Hirata, S. Aoyagi, H. Kuzuoka, and Y. Harada, Impact of seating positions on group video communication, In Proceedings of the 2008 ACM conference on Computer supported cooperative work, CSCW'08, ACM, pp.177–186 (2008).
- [32] N. Yankelovich, N. Simpson, J. Kaplan, and J. Provino, Porta-person: telepresence for the connected conference room, CHI'07 extended abstracts on Human factors in computing systems, ACM, pp.2789–2794 (2007).

Session 4:
System, Software and
Algorithm
(Chair: Takuya Yoshihiro)

Variable Coverage: A Metric to Evaluate the Exhaustiveness for Program Specifications Based on DbC

Yuko Muto[†], Yukihiro Sasaki[†], Takafumi Ohta[†], Kozo Okano[†], Shinji Kusumoto[†], and Kazuki Yoshioka[†]

[†]Graduate School of Information Science and Technology, Osaka University, Japan
 {okano, t-ohta, kusumoto}@ist.osaka-u.ac.jp

Abstract - For realizing dependability and maintainability of software, Design by Contract is one of useful notions, which utilizes constraints as contracts between the caller and the callee routines in programs. Some verifiers for programs are able to check whether given source code satisfies given constraints. It is, however, hard to measure the exhaustiveness for specification, *i.e.*, how much constraints cover ideal specification for the source code. This paper proposes Variable Coverage, a simple set of metrics to check the exhaustiveness of specification with source code for Java and other Object-Oriented programming languages. The proposed coverage observes occurrence of variables in constraints, such that the variables are also used in the target method/constructor. We applied the metrics to three concrete programs in order to evaluate that Variable Coverage is able to help to find variables which should have been referred in specifications as important variables. As a result, we found some shortage of JML annotations in target programs, which shows the usefulness of the proposed metrics.

Keywords: DbC, Coverage, Specification, Testing, Metrics

1 Introduction

Formal methods[1], mathematical techniques for the specification, development and verification of software and hardware systems, have attracted much attention because they are said to play important roles for designing software as increasing the size of software. The larger program sizes, the more frequently software testing misses corner-case. On the other hand, formal methods have potential for exhaustive checking. In the industry, some real large projects succeeded using formal methods, such as the public transportation systems[2]. Formal methods are classified into three technologies: deductive methods, model checking, model-based simulation or testing.

Design by Contract (DbC) [3] is a well-known notion to clarify the responsibility between callers and callees. Java Modeling Language (JML) [4]–[6] is a specification language for Java based on DbC. Program based on DbC can be verified with some techniques, static checking and runtime checking. For example, ESC/Java2[7] and jml4c[8] are such tools for Java. For another language, Spec# [9] is a superset of C#, and the static checker for Spec# developed by Microsoft uses Boogie[10].

It is, however, hard to determine whether the specification is well-written (exhaustive) or not. If the specification is low exhaustive, the correctness of the program is not clear. Take runtime checking as an example. A runtime checker pro-

duces a violation when source code and its specification do not meet. No violation is reported by runtime checkers if the code has no specification because there is not any constraint to check. Consequently we cannot anything about the quality of the source code.

Some papers have studied coverage metrics for hardware verification. Paper[11] summarizes some coverage metrics for simulation-based verification such as code coverage, assertion coverage. In order to generate test efficiently, Paper [12] has proposed functional coverage as the amount of control behaviors covered by the test suite using abstraction techniques. Nevertheless, there are few coverage metrics which can be applied to general purpose programming languages at the implementation level such as JML.

In this paper, we propose Variable Coverage as coverage metrics for formal specifications at the implementation level. Variable Coverage consists of the coverage for pre-condition, post-condition, assignable and invariant.

We have experimented it to apply the three kinds of programs using a prototype which measures Variable Coverage. As a result, we found some shortage of JML annotations in the target programs, which shows the usefulness of our proposed metrics.

The paper is organized as follows. Section 2 provides the definitions of some words as preliminary, and Section 3 mentions the related works. Section 4 will show our proposed method, Variable Coverage, followed by experiments and discussion in Sections 5 and 6, respectively. Finally, Section 7 concludes this paper.

2 Preliminaries

This section provides some concepts and the definition of as preliminaries.

2.1 Design by Contract

Design by Contract (DbC) is a notion proposed by Bertrand Meyer [3]. In DbC, suppliers (caller routine) and clients (callee routine) make contracts each other. Clients should satisfy the pre-conditions, and suppliers should satisfy the post-conditions under pre-conditions. This mechanism makes it easy to identify bugs.

Some programming languages support DbC as standard, others have the specification language separately from the core grammar of the language. Eiffel[13] supports DbC as standard. C# and Java have no standard contract system but some specification languages are proposed separately. Spec#[9]

is a superset of C# to describe contracts. For Java, Java Modeling Language[4] is the de-facto standard specification language.

2.2 Constraints

Pre-condition for a routine (method) is a set of boolean constraints. It should be `true` prior to the routine execution. Clients are responsible to meet pre-condition.

Post-condition for a routine is a set of boolean constraints. It should be `true` after the routine execution provided that its associate pre-condition holds. Suppliers are responsible to meet post-condition under the pre-condition.

The routine is permitted to assign values to only the variables specified in Assignable. The constrains provide to detect side effects for developers.

Invariant is a set of boolean constraints. It should be always `true`. Depending on the target of constraint, invariants are divided into class invariant and loop invariant. This paper deals with only the class invariant.

2.3 Java Modeling Language

Java Modeling Language (JML) is a specification language based on Design by Contract for Java. JML supports pre-condition, post-condition, assignable and invariant. We explain them through class `BankAccount`, an account for customer of a bank, as example.

Figure 1 is the source code of class `BankAccount` with JML.

Pre-conditions Keyword `@requires` is used to express the pre-condition. In Fig. 1, methods `withdraw` and `deposit` have pre-conditions at lines 12,13 and 20.

Post-conditions Keyword `@ensures` means the post-condition.

The constructor and methods `withdraw`, `deposit` and `getBalance` have post-conditions. Line 6 in Fig. 1 means that field `balance` is 0 after instance creation.

Assignables `@assignable` is used to express assignable.

The following `@assignable` classes, fields which can be assigned are listed. If every field is prohibited to be assigned, describe `@assignable \nothing` like line 28 in Fig. 1. Furthermore, `@pure` is equivalent to `@assignable \nothing`; this is used to make it short.

Invariants The JML description of invariants is `@invariant`. Also, if an attribute `a` with `@non_null`, it is equivalent to `@invariant a != null`. In Fig. 1, line 4 is invariant clause which means field `balance` must be 0 or more at any time.

2.4 Global Variables

Generally, the word “global variables” are not used in object-oriented programming language. In this paper, as a matter of convenience, we define global variables as follows.

```

1 public class BankAccount {
2
3     private int balance;
4     // @invariant balance >= 0;
5
6     // @ensures balance == 0;
7     // @assignable balance;
8     public BankAccount() {
9         this.balance = 0;
10    }
11
12    // @requires amount >= 0;
13    // @requires balance >= amount;
14    // @ensures balance == \old(balance) +
15        amount;
16    // @assignable balance;
17    public void withdraw(int amount) {
18        this.balance -= amount;
19    }
20
21    // @requires amount >= 0;
22    // @ensures balance == \old(balance) +
23        amount;
24    // @assignable balance;
25    public void deposit(int amount) {
26        this.balance += amount;
27    }
28
29    // @ensures \result == balance;
30    // @assignable \nothing;
31    public int getBalance() {
32        return this.balance;
33    }
34
35    // @pure
36    public void inquiry() {
37        System.out.println("Balance is " + this.
38            balance);
39    }
40 }

```

Figure 1: Source Code of Class `BankAccount` with JML

Definition 2.1 (Global Variables)

When a method m is a member of class c , a global variable g is defined as:

- g is not a member of c , and
- g is visible from m

Figure 2 shows an example of a global variable. A variable `font` of class `Config` is a global variable for method `draw`.

3 Related Work

This section introduces some works related to this paper.

3.1 Program Verification

ESC/Java[16], an Extended Static Checker for Java, is the practical usable checker among early verifiers. Currently, its successor version, ESC/Java2[17] is widely used, and it supports JML2.

Supporting the newer Java, Mobius[18] attracts rising attention as a program verification environment (PVE), including static checkers, runtime checkers and verifiers. It is provided as an Eclipse[19] plug-in. ESC/Java2 is also integrated into Mobius.

```

1 public class Config {
2     public static Font font;
3 }
4
5 public class Customer {
6     public void draw(Graphics g) {
7         g.setFont(Config.font);
8         g.drawString("An example for a global
9             variable", 10, 10);
10    }
11 }

```

Figure 2: An Example of a Global Variable

3.2 Verification Coverage

Coverage metrics for formal verification are called verification coverage in mainly the field of hardware. Verification coverage falls into two categories: syntactic coverage and semantic coverage[11]. As syntactic coverage, code coverage for model-based simulation is the metrics derived from software testing[20]. The ratio of executed code during the simulation is code coverage. As simple coverage, line coverage, the code of block without control transition.

Coverages depending on control flow graph (CFG), are branch coverage, expression coverage, path coverage.

For semantic coverage, there are functional coverage and assertion coverage. Assertion coverage is the measuring method which users determine variables which to observe. The assertion coverage measures what assertions are covered with a given set of input sequence[11].

In order to generate test suite and analyze it, paper [12] proposed functional coverage which is the amount of control behavior covered by the test suite using abstraction techniques.

3.3 Assertion Density

Assertion density is the number of assertions per line of code[21]. Without sufficient assertion density, the full benefits of assertions are not realized. Assertions must be verified, which are behaviors as design intents, *i.e.*, statements for properties.

4 Variable Coverage

This section defines Variable Coverage, our proposed method.

4.1 Motivation

Formal verification checks consistency between source code and its constructs based on Class Correctness formula. Paper[11] also states that “Measuring the exhaustiveness of a specification in formal verification has a similar flavor as measuring the exhaustiveness of the input sequences in simulation-based verification for hardware.” Applying the idea to software, the input sequences of a method/constructor correspond to variables. Consequently, we propose a coverage metric which observes variables.

4.2 Policies

We propose a set of metrics which supports these policies:

1. Our metric checks all variables as input and output. It is oriented from verification coverage.
2. Our metric is simple. The execution of measuring the coverage requires enough short time. The metric targets developers who describe assertions in JML. Our metric should be checked in short time on a frequent basis when they want to check.
3. Our metric uses only static information. Using only static information (source code and JML) without execution trace enables to measure coverage for a part of incomplete code.

4.3 Constraints Development Process with Variable Coverage

Quickly measuring Variable Coverage (VC in short) enables to high-frequently measure it. Implementators can improve constraints description by the iterative process:

Step1 Implementators describe assertions

Step2 VC is measured

Step3 Iterate Step1 if implementators find lack of their assertions

We call such an iteration “Quick VC revise.”

4.4 Definition of Variable Coverage

VC consists of four kinds of metrics, including coverage for pre-condition, post-condition, assignable and invariant. Tables 1 and 2 show VC metrics for a single constraint and multiple constraints, respectively.

4.4.1 The Coverage for Pre-conditions

Pre-conditions should check all input variables, *i.e.*, parameter of the method, attributes and global variables referred in the method. Thus, the Coverage for Pre-conditions consists of Parameters Coverage, and Referred Attributes Coverage.

Definition 4.1 (PrPC)

Let $P(m)$, and $P_{held-by-pre}(m)$ be a set of parameters defined in method m , and held by pre-condition in method m , respectively. Equation (1) defines $PrPC(m)$, Parameters Coverage for pre-conditions of method m .

$$PrPC(m) = \frac{|P_{held-by-pre}(m)|}{|P(m)|} \quad (1)$$

For Fig. 3, $|P_{held-by-pre}(m)| = |\{age\}| = 1$, and $|P(m)| = |\{name, age\}| = 2$ hold. Hence, we have $PrPC(m) = 1/2$.

Definition 4.2 (PrAC)

Let $A_{referred}(m)$, and $A_{held-by-pre}(m)$ be a set of attributes referred in method m , and held by pre-condition in method m , respectively. Equation (2) defines $PrAC(m)$, Referred Attributes Coverage for pre-conditions of method m .

$$PrAC(m) = \frac{|A_{held-by-pre}(m)|}{|A_{referred}(m)|} \quad (2)$$

Table 1: Variable Coverage (single-constraint)

Coverage Name	Constraint	Target Variables	Measuring Unit
PrPC	Pre-Condition	Parameters	Method
PrAC		Referred attributes	Method
PrGC		Referred global variables	Method
PoRC	Post-Condition	Return value	Method
PoAC		Assigned attributes	Method
PoGC		Assigned global variables	Method
AAC	Assignable	Assigned attributes	Method
IAC	Invariant	Attributes	Class

Table 2: Variable Coverage (multi-constraint)

Coverage Name	Constraint	Target Variables	Measuring Unit
PrIAC	Pre-condition + invariant	Referred attributes	Method
PoIAC	Post-condition + invariant	Assigned attributes	Method

```

1 //@ requires age >= 0;
2 // no requires holds 'name'
3 public Customer(String name, int age){
4     this.name = name;
5     this.age = age;
6 }

```

Figure 3: An Example to Explain Parameters Coverage for Pre-condition

Definition 4.3 (PrGC)

Let $G_{referred}(m)$, and $G_{held-by-pre}(m)$ be a set of global variables referred in method m , and held by pre-condition in method m , respectively. Equation (3) defines $PrGC(m)$, Referred Global Variables Coverage for pre-conditions of method m .

$$PrGC(m) = \frac{|G_{held-by-pre}(m)|}{|G_{referred}(m)|} \quad (3)$$

4.4.2 The Coverage for Post-conditions

Post-conditions observe output variables which affect the outside of method, *i.e.*, return value, attributes and global variables assigned in the method. Hence, the coverage for post-condition is composed of Return Value Coverage, Assigned Attributes Coverage and Assigned Global Variables Coverage.

Definition 4.4 (PoRC)

Equation (4) defines $PoPC(m)$, Parameters Coverage for post-conditions of method m .

$$PoRC(m) = \begin{cases} 1 & \text{(return value is held by post-condition)} \\ 0 & \text{(otherwise)} \end{cases} \quad (4)$$

Definition 4.5 (PoAC)

Let $A_{assigned}(m)$, and $A_{held-by-post}(m)$ be a set of attributes assigned in method m , and held by post-condition in method m , respectively. Equation (5) defines $PoAC(m)$, Assigned

Attributes Coverage for post-conditions of method m .

$$PoAC(m) = \frac{|A_{held-by-post}(m)|}{|A_{assigned}(m)|} \quad (5)$$

Definition 4.6 (PoGC)

Let $G_{assigned}(m)$, and $G_{held-by-post}(m)$ be a set of global variables assigned in method m , and held by post-condition in method m , respectively. Equation (6) defines $PoGC(m)$, Assigned Global Variables Coverage for post-conditions of method m .

$$PoGC(m) = \frac{|G_{held-by-post}(m)|}{|G_{assigned}(m)|} \quad (6)$$

4.4.3 The Coverage for Assignables

Constraints assignable are written on methods or constructors. Some variables are assigned in the method or constructor, among them attributes have the scope of method outside. Thus, coverage for assignable includes Assigned Attributes Coverage.

Definition 4.7 (AAC)

Let $A_{assigned}(m)$, and $A_{held-by-assign}(m)$ be a set of attributes assigned in method m , and held by assignable in method m , respectively. Equation (7) defines $AAC(m)$, Assigned Attributes Coverage for assignable of method m .

$$AAC(m) = \frac{|A_{held-by-assign}(m)|}{|A_{assigned}(m)|} \quad (7)$$

4.4.4 The Coverage for Invariants

Class invariants are described in a class. The variables owned by classes are attributes. Hence, coverage for invariants has Attributes Coverage for invariant.

Definition 4.8 (IAC)

Let $A(c)$, and $A_{held-by-inv}(c)$ be a set of attributes owned by class c , and held by invariants in class c , respectively. Equation (8) defines $IAC(c)$, Attributes Coverage for invariant of

class c .

$$IAC(c) = \frac{|A_{held-by-inv}(c)|}{|A(c)|} \quad (8)$$

4.4.5 The Coverage for Pre-conditions and Invariants

Definition 4.9 (PrIAC)

Let assume that Class c owns method m . Also let $A_{referred}(m)$, $A_{hold-by-pre}(m)$, and $A_{hold-by-inv}(c)$ be a set of attributes referred in method m , held by pre-condition in method m , and held by invariants in class c , respectively. Equation (9) defines $PrIAC(m)$, Referred Attributes Coverage for pre-conditions and invariants of method m .

$$PrIAC(m) = \frac{PrIACNR(m)}{|A_{referred}(m)|} \quad (9)$$

, where $PrIACNR(m) = |A_{referred}(m) \cap (A_{held-by-pre}(m) \cup A_{held-by-inv}(c))|$

4.4.6 The Coverage for Post-conditions and Invariants

Definition 4.10 (PoIAC)

Let assume that Class c owns method m . Let $A_{assigned}(m)$, $A_{hold-by-post}(m)$, and $A_{hold-by-inv}(c)$ be a set of attributes referred in method m , held by post-condition in method m , and held by invariants in class c , respectively. Equation (10) defines $PoIAC(m)$, Assigned Attributes Coverage for post-conditions and invariants of method m .

$$PoIAC(m) = \frac{PoIACNR(m)}{|A_{assigned}(m)|} \quad (10)$$

, where $PoIACNR(m) = |A_{assigned}(m) \cap (A_{held-by-post}(m) \cup A_{held-by-inv}(c))|$

4.4.7 Ignored Variables

Constants are ignored from measuring the coverage because such variables do not affect on communication among methods. For example, in Java, the variables decorated by `final` modifier are ignored.

5 Evaluation

This section gives the experimental evaluation and the results.

5.1 Overview

We performed experimentation using our prototype tool in order to evaluate our proposed coverage metrics. We measured (1) execution times, and (2) the numeric results of our proposed coverage. Here is the experimental environment; Machine is HP Z800 Workstation (Xeon E5607 dual core 2.27GHz, 2.26GHz and main memory 32GB); It was installed Windows 7 Professional for 64bit with Service Pack 1 and Java Version 1.7.

5.2 Target Programs

We apply our approach to three programs: Warehouse Management Program (WMP)[22], HealthCard (HC)[23], [24], and Syllabus Management System for a university (SMS). Table 3 summerizes the target programs including the size of programs and available assertion types which the program has.

Table 3: Target Programs

Target Program	N	Available JML Assertions
WMP	53	requires,ensures, assignable,invariant
HC	197	requires,ensures,assignable
SMS	562	requires,ensures

\bar{N} = The number of target methods and constructors

WMP is developed by an ex-member of our research group. This program has all of `requires`, `ensures`, `assignables` and `invariants`, and they are whole passed by the static checker, ESC/Java2.

HC is a medical appointment application which is written as a master thesis work by Joao Pestana Ricardo Rodrigues from University of Madeira. It is based on JavaCard, the platform of IC card devices. In general, the embeded systems need more strictly quality because it is hard to update their software. The HealthCard has two versions: running version and JML version. We utilize JML version as experimental target because JML version contains more JML description than running version. HealthCard program has no `@invariant` in JML because `model` is used instead of `@invariant`. Thus, in this evaluation, Attributes Coverage for Invariant are not measured.

SMS is implemented in Java by a software company as an educational resource for IT Specialist Program Initiative for Reality-based Advanced Learning (IT Spiral), a national educational project leading by MEXT. Members of our research group added only preconditions and postconditions in JML into the system, and the system produces no violations by `jml4c`, a runtime checker.

We add the standard libraries (e.g., `java.lang.Object`) with JML descriptions[17] into target programs. Thus, into the class which inherits a class or implements a interface, the contracts of its superclass or interface are added. For example, the contracts of `java.lang.Object#toString()` are added into all methods `toString()`. As well, the results of coverage do not include the methods of the standard libraries. Furthermore, we excluded abstract classes, interfaces, test classes and the main method because they should not have necessarily contracts.

5.3 Results of Execution Times

Table 4 shows the results of execution times. We measured three execution times for each program; it shows the average of them.

Table 4: Execution Times

Target Program	Execution Time
WMP	9.3 sec
HC	16.0 sec
SMS	14.0 sec

5.4 Results of Variable Coverage

Tables 5, 6, and 7 show the results of coverages for pre-conditions, for post-conditions, and for assignable, respectively.

Table 5: Results of Coverage for Pre-conditions

Target Program	PrPC	PrAC	PrIAC
WMP	99.17%	9.09%	96.97%
HC	79.22%	46.24%	NA
SMS	41.82%	2.77%	NA

Table 6: Results of Coverage for Post-conditions

Target Program	PoRC	PoAC	PoIAC
WMP	100.00%	94.12%	100.00%
HC	84.11%	48.39%	NA
SMS	99.68%	99.38%	NA

Table 8 shows the results of coverage of invariant for Warehouse Management System.

6 Discussion

This section discusses the experimental results and the threats to validity.

6.1 Warehouse Management Program

The following method does not cover Parameter Coverage for pre-conditions:

```
StockManagement.Request#
Request(java.lang.String, int,
StockManagement.Customer, java.util.Date,
byte).
```

We found that parameter `rqst` is not covered by `requires` in source code of the constructor `Request`. The byte-type parameter `rqst` means the request state instead of Enum, as `SHORTAGE=0`, `SATISFYED=1`, `DELIVERED=2`, `WAIT=3`. Therefore, constraints of class `Request` in JML lack because its attribute `rqst` must be any of 0 to 3.

Table 6 shows that every return value is held by its post-conditions. No problem was found when we read the source code and JML.

For Assigned Attributes Coverage for Post-condition, the following method does not cover it:

```
StockManagement.ReceptionDesk#
ReceptionDesk().
```

Developers who described the source code and JML seemed to recognize the shortage of post-condition, because there is

Table 7: Results of Coverage for Assignable

Target Program	AAC
WMP	100.00%
HC	41.94%
SMS	NA

Table 8: Results of Coverage for Invariant (Warehouse Management Program)

Class Name	P	IAC
ContainerItem	3 / 3	100.00%
Customer	3 / 3	100.00%
Item	2 / 2	100.00%
ReceptionDesk	2 / 2	100.00%
Request	4 / 6	66.67%
StockState	NA	NA
Storage	3 / 3	100.00%

P=The number of attributes held by invariants / the Number of attributes

a comment “ensures are included in invariants” in the source code (Fig. 4).

```
1 //ensures are included in invariants.
2 //@ public behavior
3 //@ assignable requestList, storage;
4 public ReceptionDesk() {
5     requestList = new LinkedList();
6     storage = new Storage();
7 }
```

Figure 4: Constructor `ReceptionDesk` which is Not Covered by Post-conditions

In the source code of class `ReceptionDesk` (Fig. 5), attributes `requestList` and `storage` are held by invariants.

Also, the result of Assigned Attributes Coverage for Post-condition and Invariant is 100%. Even if Assigned Attributes Coverage for Post-condition is low, we can conclude that the source code does not have the problem because the value of Assigned Attributes Coverage for Post-condition is high. Hence, VC helps us to clarify that source code has no problem.

Like the case of class `ReceptionDesk`, it is hard to know the reason why post-conditions are omitted in a general case. One solution is the designer should describe a comment or some keyword when the post-conditions are included in the class invariants.

Table 7 shows that all assigned attributes are held by assignables. Therefore, we can see that every assignable is described rightly in Warehouse Management Program.

Table 8 shows that Attributes Coverage for Invariant of most of classes have 100%, but the coverage of class `Request` is 66%. Class `Request` has six attributes but the two of all are not held by invariant constraints. We found that attributes `deliveringDate` and `requestState` in class `Request`, are the cause. `deliveringDate` is defined as `java.util.Date` type field which means the date of de-

```

public class ReceptionDesk {
  private /*@ spec_public non_null @*/ List
    requestList;
  private /*@ spec_public non_null @*/ Storage
    storage;
  /*@ public invariant \typeof(requestList) ==
    \type(Request);
    ...
}

```

Figure 5: Invariants in Class ReceptionDesk

Table 9: Extracted Results of Coverage for HealthCard

T	N	PrPC	PrAC	PoRC	PoAC	AAC
(1)	197	79.22 %	46.24 %	84.11%	48.39%	41.94 %
(2)	38	82.61%	42.86 %	88.89%	NA	NA

T=The Type of Targets

N=The Number of Targets

(1):All methods and constructors

(2):Except for constructors, setters and getters

living. Any field of type `java.util.Date` except for `deliveringDate` in class `Request` has a constraint “the field is not null.” Thus, the implementor has no idea about the constraints of `deliveringDate` because `deliveringDate` can be null before delivering. The same is true respect to field `requestState`. Figure 6 shows our suggested revised version of constraints for them based on the results.

```

1 public class Request implements Comparable {
2   private /*@ spec_public non_null @*/ Date
3     receptionDate;
4   private /*@ spec_public non_null @*/ String
5     itemName;
6   private /*@ spec_public @*/ int amount;
7   private /*@ spec_public non_null @*/
8     Customer customer;
9
10  private byte requestState;
11  private Date deliveringDate;
12  /*@invariant
13  (requestState != delivered &&
14    deliveringDate == null) ||
15  (requestState == delivered &&
16    deliveringDate != null);
17  ...
18 }

```

Figure 6: Class Request with JML We Suggest

6.2 HealthCard

From the manual inspection we conclude that the JML assertion for `HealthCard` is described in the following way. No constructors has JML description because JML description is on interface. Setters and getters have no JML description. We discuss constructors, setters and getters later. Table 9 includes the results of `HealthCard` except for constructors, setters and getters.

According to Table 9, the following methods have no pre-condition with their parameters though they are setters/getters

nor constructors:

- `commons.CardUtil#byte[] clone(byte[])`
- `commons.CardUtil#void cleanField(byte[])`
- `commons.CardUtil`
#boolean `validateObjectArrayPosition`
(`java.lang.Object[]`, short)
- `commons.CardUtil`
#short `countNotNullObjects`
(`java.lang.Object[]`)

The parameters of the methods are array type, and any caller or any callee dose not guarantee that each of the parameters is not null. We found the shortage of JML descriptions by applying Variable Coverage. In addition, the methods do not check whether their parametera are null or not in their body. `NullPointerException` is thrown when the parameter array is null. It shows that these methods have potential bugs.

Also, there is a method with comments in natural language instead of constraints in JML. Figure 7 shows the source code of method

`validateObjectArrayPosition` of class `CardUtil`. Line 1 in the figure indicates that the developers know the lack of JML descriptions. We consider, as future work, that it is possible to infer contracts from useful comments.

```

1 //Returns false if position points to a null
2   value or if position is out of bounds.
3 //@ assignable \nothing;
4 public /*@ pure @*/ static boolean
5   validateObjectArrayPosition (Object[]
6   array, short position) {
7   if(position < 0 || position >=
8     countNotNullObjects(array))
9     return false;
10  else
11    return true;
12 }

```

Figure 7: Comments Instead of Contracts

For Referred Attributes Coverage for Pre-condition, the results of 23 methods are not full coverage. For methods `toString` occupying 8 of 23, they are eliminated from the results because their source code have the comments, “Testing code.”

One behalf of other 15 methods, we explain a method `validateAllergyPosition`. It does nothing other than calling utility method `validateObjectArrayPosition` of class `CardUtil` (Fig. 8).

```

1 public boolean validateAllergyPosition(short
2   position){
3   return CardUtil.validateObjectArrayPosition(
4     this.allergies, position);
5 }

```

Figure 8: Source Code of Method `validateAllergyPosition`

It is preferable that contract violations are produced at previous step than later because it makes easy to identify bugs. Thus, methods `validateAllergyPosition` and `validateVaccinePosition` should be written more JML descriptions.

For Return Value Coverage for Post-condition, in analogy with pre-conditions, the following methods have no post-conditions in spite they are setters/getters nor constructors:

- `commons.CardUtil#byte[] clone(byte[])`
- `commons.CardUtil`
#short countNotNullObjects
(`java.lang.Object[]`)
- `commons.CardUtil`
#boolean validateObjectArrayPosition
(`java.lang.Object[]`, short)

The JML descriptions of the methods can be improved. For method `clone`, we suggest the post-conditions `@ensures \result != null`. Also, we have the idea like Figure 9 for method `validateObjectArrayPosition`, from its comment “//Returns false if position points to a null value or if position is out of bounds.”:

```
/*@ ensures
  (\result == false) ==>
  (array == null ||
   position <= 0 || position >=
    countNotNullObjects(array))
  @*/
```

Figure 9: Post-condition of Method `validateObjectArrayPosition` which We Recommend

For method `countNotNullObjects`, we suggest `@ensures \result >= 0;`.

About Assigned Attributes Coverage for Post-condition, the result is not available because there are no methods which assign to the attributes.

There are no methods which assign the attributes except for constructors, setters and getters.

In general, constructors and setters tend to change the attributes. Although every getter does not change the attributes, its return value is used by other methods. In order to guarantee the behavior of the class, constructors, setters and getters should have JML descriptions.

We recommend for developers to describe the JML description of constructors, setters and getters like Figure 10. To setters, developers should write pre-condition which means that parameters equals attributes assigned. To getters, developers should write post-condition which means that return value equals attributes returned.

6.3 Syllabus Management System

The parameters of 207 methods are not held by pre-conditions; 144 of them are setters, and 63 are others. As an instance of setters, Figure 11 shows the source code of method `setJugyouKamoku` of class `JikanwariJugyouKamokuDTO`.

```
public class Person {
    private String name;

    //@requires name != null;
    //@ensures this.name == name;
    public Person(String name) {
        this.name = name;
    }

    //@requires name != null;
    //@ensures this.name == name;
    public void setName(String name) {
        this.name = name;
    }

    //@ensures \result == this.name;
    //@assignable nothing;
    public String getName() {
        return this.name;
    }
}
```

Figure 10: Source Code with JML of Setter and Getter We Recommend

When parameter `jugyouKamoku` is null, the attribute `jugyouKamoku` is set to null.

If method `setJugyouKamoku` are called again, the null reference is occurred at line 2. Thus, pre-condition should have the constraints for parameter `jugyouKamoku` which means `jugyouKamoku != null`.

```
1 //@ ensures this.jugyouKamoku.equals(
   jugyouKamoku);
2 public void setJugyouKamoku(final JugyouKamoku
   jugyouKamoku) {
3     this.jugyouKamoku = jugyouKamoku;
4 }
```

Figure 11: An Example for Setter of Syllabus Management System

Only the following method does not have full coverage for Return Value Coverage for Post-Condition:

```
service.UserServiceImpl#
boolean authenticate(java.lang.String,
java.lang.String, entity.UserKubun)
```

The method `authenticate` of class `UserServiceImpl` returns true or false depending on its parameters. We found no post-condition in its source code; It is hard to distinguish from forgetting constraints. Therefore, for such a method, we recommend to write explicitly these contract to alternate from oversights:

```
ensures \result == true|false;
```

For Assigned Attributes Coverage for Post-condition, the result of the below method is not held by post-conditions:

```
entity.Soshiki # void add(entity.Soshiki)
```

Figure 12 shows the source code of the method `add` of class `Soshiki`. Post-condition at Line 3 calls getter method `getKaiSoshiki`. From The source code of the getter (Figure 13), the getter just returns the attribute `kaiShoshiki` without changing it. We recommend to describe

ensures this.kaiSoshiki.contains(soshiki);
instead of line 3.

```

1  //@ requires soshiki != null;
2  //@ ensures this.getKaiSoshiki().contains(
   soshiki);
3  public void add(final Soshiki soshiki) {
4      if (getKaiSoshiki() == null) {
5          this.kaiSoshiki = new LinkedHashSet<
           Soshiki>();
6      }
7      soshiki.setJouiSoshiki(this);
8      getKaiSoshiki().add(soshiki);
9  }

```

Figure 12: Source Code of Method add of Class Soshiki

```

1  //@ ensures (this.kaiSoshiki != null) ? (this.
   kaiSoshiki.size() == \result.size()) && (
   \forall s Soshiki s; this.kaiSoshiki.
   contains(s); \result.contains(s)) : \
   result == null;
2  // anotation OneToMany(cascade = CascadeType.
   ALL, targetEntity = Soshiki.class,
   mappedBy = "jouiSoshiki")
3  public Set<Soshiki> getKaiSoshiki() {
4      return this.kaiSoshiki;
5  }

```

Figure 13: Source Code of Method getKaiSoshiki of Class Soshiki

Calling the setter of the attribute in the methods is the same as assigning the attribute. For example, line 5 at Figure 14 is equivalent to assigning the attribute SESSION. Assigned Attributes Coverage should be extended to target calling the setter of the attribute additionally.

```

1  public static Session currentSession() {
2      Session s = SESSION.get();
3      if (s == null) {
4          s = SESSION_FACTORY.openSession();
5          SESSION.set(s);
6      }
7      return s;
8  }

```

Figure 14: Example for the Unmonitored Case of Assigning to An Attribute

7 Conclusion

This paper proposed Variable Coverage, a set of metrics for the exhaustiveness of specification with source code based on Design by Contract. Our proposed coverage observes some variables depending on constraints. We applied our approach to three programs in order to evaluate that Variable Coverage is able to help to find variables which should have been referred in specifications as important variables. As a result, we found some shortage of JML annotations in target programs, which shows the usefulness of our proposed metrics.

Future work includes to infer the constraints to describe. The first idea is suggesting constraints to describe from the comments in the source code. The second idea is using the modifiers of method; static methods should not have assignable clause except for static variables, which means no attributes are permitted to assign, because static methods do not change the internal state (*i.e.*, attributes). Such a modifier helps to generate helpful assertions.

Acknowledgments

This work is being conducted as a part of Grant-in-Aid for Scientific Research C(21500036) and S(25220003).

REFERENCES

- [1] Edmund M. Clarke and Jeannette M. Wing. Formal Methods: State of the Art and Future Directions. *ACM Computing Surveys*, 28(4):626–643, December 1996.
- [2] Jean-Raymond Abrial. Formal Methods in Industry. In *Proceeding of the 28th international conference on Software engineering - ICSE '06*, page 761, New York, New York, USA, May 2006. ACM Press.
- [3] Bertrand Meyer. Applying ‘Design by Contract’. *IEEE Computer*, 25(10):40–51, 1992.
- [4] Gary T Leavens, Albert L Baker, and Clyde Ruby. JML: A Notation for Detailed Design. *Behavioral Specifications of Businesses and Systems*, pages 175–188, 1999.
- [5] Patrice Chalin, Perry R. James, and George Karabotsos. JML4: Towards an Industrial Grade IVE for Java and Next Generation Research Platform for JML. In Nataraajan Shankar and Jim Woodcock, editors, *Proceeding VSTTE '08 Proceedings of the 2nd international conference on Verified Software: Theories, Tools, Experiments*, volume 5295 of *Lecture Notes in Computer Science*, pages 70–83, Berlin, Heidelberg, October 2008. Springer.
- [6] David R. Cok. OpenJML: JML for Java 7 by extending OpenJDK. In *Proceeding NFM'11 Proceedings of the Third international conference on NASA Formal methods*, pages 472–479, April 2011.
- [7] David R. Cok and Joseph R Kiniry. ESC/Java2: Uniting ESC/Java and JML. In *International Workshop on Construction and Analysis of Safe Secure and Interoperable Smart Devices CASSIS 2004*, volume 3362 of *Lecture Notes in Computer Science*, pages 108–128. Springer, 2004.
- [8] Amritam Sarcar. A New Eclipse-Based JML Compiler Built Using AST Merging. In *2010 Second World Congress on Software Engineering*, pages 287–292. IEEE, December 2010.
- [9] Mike Barnett, K. Rustan M. Leino, and Schulte. The Spec# Programming System: An Overview. In Gilles Barthe, Lilian Burdy, Marieke Huisman, Jean-Louis Lanet, and Traian Muntean, editors, *Construction and Analysis of Safe, Secure, and Interoperable Smart Devices*, volume 3362 of *Lecture Notes in Computer Science*, pages 49–69, Berlin, Heidelberg, 2005. Springer.

- [10] Mike Barnett, Bor-Yuh Evan Chang, Robert DeLine, Bart Jacobs, and K. Rustan M. Leino. Boogie: A Modular Reusable Verifier for Object-Oriented Programs. In *4th International Symposium, FMCO 2005*, volume 4111 of *Lecture Notes in Computer Science*, pages 364–387, 2006.
- [11] Hana Chockler, Orna Kupferman, and Moshe Vardi. Coverage Metrics for Formal Verification. *International Journal on Software Tools for Technology Transfer*, 8(4-5):373–386, April 2006.
- [12] Dinos Moundanos, Jacob A. Abraham, and Yatin V. Hoskote. Abstraction Techniques for Validation Coverage Analysis and Test Generation. *IEEE Transactions on Computers*, 47(1):2–14, 1998.
- [13] Bertrand Meyer. *Eiffel : The Language (Prentice Hall Object-Oriented Series)*. Prentice Hall, 1991.
- [14] Charles Antony Richard Hoare. An Axiomatic Basis for Computer Programming. *Communications of the ACM*, 12(10):576–580, October 1969.
- [15] Bertrand Meyer. *Object-Oriented Software Construction (2nd Edition)*. Prentice Hall, 2 edition, 2000.
- [16] Cormac Flanagan, K. Rustan M. Leino, Mark Lillibridge, Greg Nelson, James B. Saxe, and Raymie Stata. Extended Static Checking for Java. *ACM SIGPLAN Notices*, 37(5):234, May 2002.
- [17] KindSoftware. ESC/Java2.
- [18] Joseph Kiniry, Patrice Chalin, Clément Hurlin, Bertrand Meyer, and Jim Woodcock. Integrating Static Checking and Interactive Verification: Supporting Multiple Theories and Provers in Verification. In Bertrand Meyer and Jim Woodcock, editors, *VERIFIED SOFTWARE: THEORIES, TOOLS, EXPERIMENTS*, volume 4171 of *Lecture Notes in Computer Science*, pages 153–160. Springer, Berlin, Heidelberg, 2008.
- [19] Eclipse Foundation. Eclipse.
- [20] Serdar Tasiran and Kurt Keutzer. Coverage Metrics for Functional Validation of Hardware Designs. *IEEE Design & Test of Computers*, 18(4):36–45, 2001.
- [21] Harry Foster, David Lacey, and Adam Krolnik. *Assertion-Based Design*. Springer, second edition, May 2004.
- [22] Masayuki Owashi, Kozo Okano, and Shinji Kusumoto. Design of Warehouse Management Program in JML and Its Verification with ESC/Java2 (in Japanese). *The Transactions of the Institute of Electronics, Information and Communication Engineers D*, 91(11):2719–2720, 2008.
- [23] Ricardo Miguel Soares Rodrigues. JML-Based formal development of a Java card application for managing medical appointments. *University of Madeira*, 2009.
- [24] Ricardo Miguel Soares Rodrigues. HealthCard.

Development of Teaching Materials for Computer Programming using a Robot Remotely Controlled by a PC through Wireless Communication

Toshihiro Shikama*

*Fukui University of Technology
shikama@fukui-ut.ac.jp

Abstract - We developed teaching materials for students to increase their interest in computer programming. We employed a robot specified by ET Robocon (Embedded Technology Software Design Robot Contest). Although the robot in ET Robocon is controlled by a program running in the robot itself, a program written by a student runs on a separate PC and also controls the robot through wireless communication via Bluetooth. As for the programming language that students learn, we selected Python because of its simplicity and similarity with object-oriented programming. A student can start programming simple sequential control of the robot and extend it to programming that realizes line tracing.

Keywords: Embedded Systems, ET Robocon, Python, Teaching Materials, Line Tracing

1 INTRODUCTION

This paper reports the development of teaching materials (sometimes shortened to “materials” in this paper) for computer programming. When a student is learning computer programming, the initial stage is important. The student generally takes a long time to become familiar with abstract programming concepts such as data types, structures, and classes of object-oriented programming. These concepts are separate from physical instances and difficult to learn. If we educate students under the false assumption that they will easily understand these abstract concepts, the students may abandon learning because of a loss of interest.

We participated in the ET Robocon (Embedded Technology Software Design Robot Contest) so that students could learn embedded systems [1]. In this contest, students analyze and design a computer program using UML (Unified Modeling Language) to control the robot, which contains strictly defined hardware with no modifications allowed. We observed that students who participated in this contest tended to become enthusiastic about computer programming. From this experience, we expect that more students will show an interest in computer programming if we incorporate the robot into the programming education. Based on this motivation, we developed teaching materials for computer programming using the robot defined by ET Robocon.¹

¹ The work reported in the paper was supported by the Special Research Grant-in-Aid of Fukui University of Technology.

2 OUTLINE OF ET ROBOCON AND OBJECTIVES OF THIS WORK

2.1 ET Robocon

The objective of ET Robocon is to improve the capability of software technology for embedded systems. This contest uses control software targeting a two-wheel self-balancing robot using LEGO MINDSTORMS NXT [1]. Figure 1 shows the appearance of the robot and its components. The robot consists of an ultrasonic sensor, a gyro sensor, a light sensor, and three motors for the right wheel, left wheel, and tail. It is equipped with a 32-bit ARM7 microprocessor. Students develop control programs that enable the robot to autonomously trace a line around a specified course. Figure 2 shows a photo of the ET Robocon 2011 course.

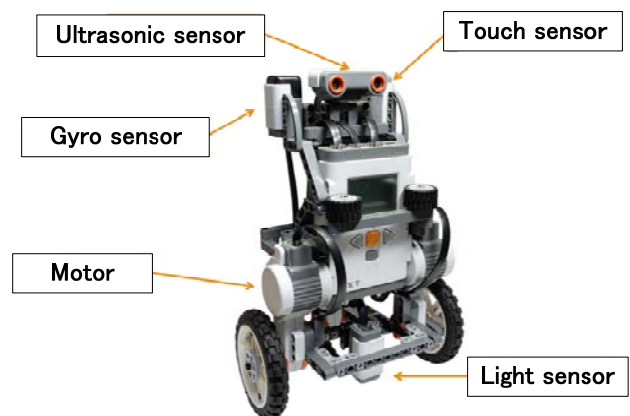


Figure 1: The robot and its components



Figure 2: ET Robocon 2011 course

The robot runs along the black lines drawn over white areas, which are surrounded by green “ground”. Students are required to develop a control program that makes the robot run along the black line at high speed. All teams in the contest use the same robot, which has a limited number of sensors (an ultrasonic sensor, a gyro sensor, and a light sensor). The ET Robocon consists of two parts: modeling and a time trial. The modeling part is a competition of the UML modeling skill used in developing the program, and the time trial part is a run-time competition of the robot. The total score is determined from the results of the two parts.

2.2 Objectives of this work

The intention of this work is to develop teaching materials for computer programming and to promote the enjoyment of computer programming for beginner students. We are aiming at the following goals:

- 1) The control of the robot is realized by a simple program (i.e., a small number of program steps).
- 2) A student can start programming without understanding abstract programming concepts.
- 3) The basics of programming skills, such as conditional branches, loops, functions, etc., can be studied through the developed materials.
- 4) The materials can be applied to the education of object-oriented programming and multi-thread programming.

3 THE BASIC ARCHITECTURE

3.1 Outline of the materials

In ET Robocon, a control program is written in the C or C++ language and the program is compiled to produce a binary file which is loaded into the robot through a USB interface. After the program is invoked, the robot is autonomously controlled by the program. Although this scheme can enable accurate and efficient control of the robot, debugging is limited since the robot has only a small display to show internal information and status. Another problem is the time and work required for students; each time a program is modified, the students must compile, link, and download the binary file through the USB interface. After considering these drawbacks, we apply the following scheme for a program running on a PC to control the robot.

- 1) A fixed control program is pre-loaded into the robot. This program executes basic commands from the PC. Students do not modify the program in the robot.
- 2) The basic commands are sent from the PC through wireless communication via Bluetooth.
- 3) Control of the robot is achieved by the program running in the PC. This program describes combinations of basic commands.
- 4) As the programming language, Python [3] is selected for the program in the PC.

3.2 Adoption of Python

As mentioned above, we adopted Python as the programming language for students to learn. Python is an object-oriented scripting language and has the following features:

- Since a program can be executed without compiling, a student can modify his program easily and test it quickly.
- Python is well defined and easy for beginners to learn.
- Python is used globally.
- A student can learn object-oriented programming easily by Python.
- Python is available at no cost and is supported by multiple platforms, including Windows and Linux.
- Because indents are mandatory in Python, a program can generally be read easily. In addition, differences in the programming style between students are small.

Adoption of Python has the following drawbacks:

- Python has a compatibility problem between versions 2 and 3.
- Performance of the program is slow because it is a scripting language.
- Indentation is employed for identifying program blocks; this programming style is different from other languages, such as C and C++.

For compatibility between Python versions, we use version 2. Our focus is on the educational aspect, and so we do not seek to realize high running speed of the robot. Concerning the indentation used in Python, we think that this is not a serious problem for students, who will study the C or C++ programming language after learning Python.

3.3 Configuration of the materials

Figure 3 shows the total configuration of the developed materials. The robot and the PC are connected through wireless communication via Bluetooth. The program running in the PC controls the robot remotely.

The Python program running in the PC sends a command through wireless communication; the robot moves forward or makes a left or a right turn by following the program commands. The running speed of the robot is also controlled by the program.

A fixed program running on the ARM 7 microprocessor inside the robot controls the movement of the robot; students do not modify this program, which is developed in the C++ language and runs on the Real-Time Operating System *nxTOSEK* [4]. This program performs the following functions:

- Controls the posture of the robot
- Receives commands through Bluetooth and interprets them
- Executes commands from the PC
- Sends log data to the PC every 40 ms

The Python module, which was developed for this setup, calculates data to get information about the robot, including the X- and Y-coordinates of the robot, the angle of the robot, and the total running distance from the start point.

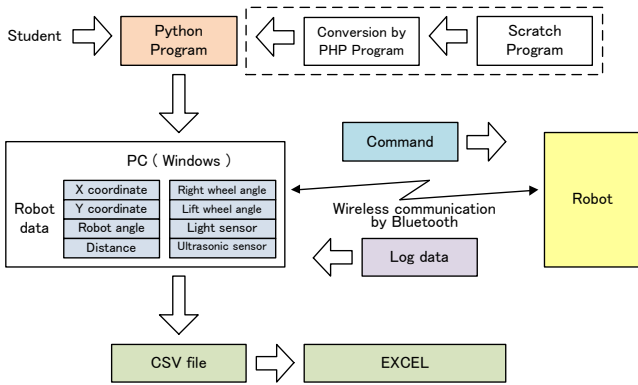


Figure 3: Configuration of the developed materials

When students write programs, they can use variables concerning these data by importing the Python module. The module also provides a log file, including all log data, in the CSV format. Using EXCEL, students can analyze the log file to obtain, for example, a trace of the robot.

Table 1 summarizes the basic functions and commands that students can use in their Python programs. For simplification, the specification commands are limited to one character, whereas extended commands consisting of multiple characters are also provided for future use.

The variable “bt” is the object to control the robot. At the head of a program, the object is generated from the defined class “nxt_bluetooth” as follows:

```
from nxt_bluetooth import nxt_bt
bt = nxt_bt("00:16:53:0c:82:39", 0).
```

The first line imports the class “nxt_bt” from the module “nxt_bluetooth”. This mandatory class was developed for the materials used. The second line generates the object “bt”, where the parameter “00:16:53:0c:82:39” is an example of the MAC address of the robot. Bluetooth employs a 48-bit MAC address, which is the same as that of LANs. As the “nxt_bt” class hides the Bluetooth communications and updates of the robot variables from students, the students can develop their programs without knowing the internal details.

Table 1: Basic functions and commands

Function	Command	Meaning
bt.send (character)	“f”	Move forward
	“r”	Make a right turn
	“l”	Make a left turn
	“b”	Move backward
	“0”-“9”	Set speed
bt.swait (seconds)	-	Wait specified seconds

Table 2 summarizes the variables of the robot status and sensors. The values of these variables are updated every 40 ms, and so students can use these variables to control the robot. For example, students can quantitatively control the robot, such as moving it forward 500 mm or making a 90-degree left turn.

Table 2: Variables of robot status and sensors

Variable	Meaning
bt.x	X-coordinate of the robot
bt.y	Y-coordinate of the robot
bt.angle	Angle of the robot
bt.distance	Total running distance from the start point
bt.diff_light	Value of the light sensor

4 INTERNAL REALIZATION SCHEME OF THE MODULE

Students import the module “nxt_bluetooth” at the head of their program. The class “nxt_bt”, which has been developed in the materials, is included in this module. This class has the functions of communication via Bluetooth, processing log data, and synchronization between the robot and the program running in the PC. As the details are hidden, students have to be aware of only the MAC address of the robot.

When we consider the implementation of the Python class, it is natural to use two separate threads for sending and receiving functions. However, to simplify the implementation, we realize the functions by a single thread, since the robot sends log data every 40 ms. We can eliminate the complexity of the multi-threading and extend the materials to realize the control of multiple robots simply by multi-threading.

Inside the nxt_bt class, sending and receiving functions via Bluetooth are realized by importing the Bluetooth module. This module is provided by python-bluez [5], which is a wrapper function that enables Python to use BlueZ [6]. The robot employs the virtual serial port communication by the serial port profile (SPP) of Bluetooth. BlueZ supports this profile. Although python-bluez is for Linux, PyBluez is also available for the Windows environment. This means that the materials can be used on both platforms, if Python is installed.

Each communication via Bluetooth is initiated by generating a socket with the required parameters and connecting it as follows, where “bt_addr” holds the character string of the MAC address.

```
self.etrobo_address = bt_addr
self.port = 1
self.sock = BluetoothSocket( RFCOMM )
try:
    self.sock.connect((self.etrobo_address, self.port))
except IOError:
    print "Robot is not invoked."
    sys.exit()
print "connected address = ", self.etrobo_address
```

After the socket has been connected, the program enters a wait state, if it calls the receive() function. As we mentioned before, since a single thread performs both sending and receiving, the program has to call the receive() function to enter the receive wait state after its process has completed. This is actually done by calling the wait() or swait() functions. The wait() function specifies a number of 40 ms

units as the wait duration, while the `swait()` function specifies the wait time in seconds.

```
def wait(self, n):
    self.i = 0
    while self.i < n:
        self.receive()
        self.i = self.i + 1

def swait(self, time):
    self.n = time // 0.04
    self.wait(self.n)
```

In the `wait()` function, the program waits for the receiving data by the `self.receive()` function. Since the robot sends log data periodically, completion of the receive occurs within 40 ms. The initial part of the `receive()` function executes the following code.

```
def receive(self):
    self.starttime = time.time()
    self.data = self.sock.recv(34)
    if len(self.data) != 34:
        print "receive byte length =", len(self.data)
        sys.exit()
    self.udata = unpack('<2BI2bH3i4hi', self.data)
```

The second line records the receive time of the log data, and then the third line extracts the received data. As the length of the log data is fixed in units of 34 bytes, the log data is divided into pre-defined formats and stored, if the data length is normal (unpack process). The unpacked data is used to calculate X- and Y-coordinates and the angle of the robot. These calculated values are stored in the Python variables, which students can use in their programs.

As mentioned above, one of the features of the materials described in this paper is that the module and programs, including the one used inside the robot, are completely open (i.e., white box). We are able to customize the robot itself and the Python module for future requests from students as well as teachers.

5 EXAMPLES OF PROGRAMS USING THE DEVELOPED MATERIALS

To explain the use of the developed materials, it is appropriate to show some program samples. We will show examples of a simple sequence control, usage of loops, usage of functions, and a simple line trace in the following.

5.1 Example 1

The program shown in Figure 4 is a basic program that controls the robot sequentially. After the program is invoked, the robot moves forward for 2 seconds, turns right for 2 seconds, moves forward for 2 seconds, turns left for 2 seconds, and then stops. Each time a command is sent, the next command is issued after the time specified by the `swait()` function. Since the program is written in Python, the program file has the extension `“py”`. If the name of the

program is `“sample1.py”`, the program is invoked by typing the following command in a terminal window.

```
python sample1.py
```

```
from nxt_bluetooth import nxt_bt
bt = nxt_bt("00:16:53:0c:48:1e", 0)

print "START"
bt.send( "3" )
bt.send( "f" )
bt.swait(2)
bt.send( "r" )
bt.swait(2)
bt.send( "f" )
bt.swait(2)
bt.send( "l" )
bt.swait(2)
bt.send( "0" )
print "END"
```

Figure 4: Example 1—sequential control

5.2 Example 2

The program shown in Figure 5 uses a `“while”` loop to check the value of a variable repeatedly. The execution leaves the loop if the variable takes a specific value. Here, the robot moves forward 500 mm (50 cm), then it makes a 180-degree left turn. After this it moves 50 cm forward again and then stops. By using the variable `“bt.distance”` that indicates the total distance from the start point and the `“while”` loop, the program can control the moving distance quantitatively. When the robot makes a turn, the angle of the robot can also be controlled in the same manner.

```
from nxt_bluetooth import nxt_bt
bt = nxt_bt("00:16:53:0c:48:1e", 0)

print "START"
bt.send( "f" )
bt.send( "3" )
while bt.distance < 500:
    bt.swait(0.04)
bt.send( "l" )
while bt.angle < 180:
    bt.swait(0.04)
bt.send( "f" )
target_dist = bt.distance + 500
while bt.distance < target_dist:
    bt.swait(0.04)
bt.send( "0" )
print "END"
```

Figure 5: Example 2—while loops

5.3 Example 3

The program shown in Figure 6 defines functions concerning an advance and a left turn. Each function takes a

parameter: a distance or an angle. This program makes the robot move forward 300 mm, make a 180-degree left turn, and then move forward 300 mm. The program repeats these actions four times by using the “for” loop.

```

from nxt_bluetooth import nxt_bt
bt = nxt_bt("00:16:53:0c:48:1e", 0)

def forward(distance):
    t_distance = bt.distance + distance
    bt.send( "f" )
    while bt.distance < t_distance:
        bt.swait(0.04)

def left_turn(angle):
    t_angle = bt.angle + angle
    bt.send( "l" )
    while bt.angle < t_angle:
        bt.swait(0.04)

print "START"
bt.send( "3" )
bt.swait(0.04)
for var in range(0, 4):
    forward(300)
    left_turn(180)

bt.send( "0" )
print "END"

```

Figure 6: Example 3—“for” loop

5.4 Value of the light sensor

Figure 7 shows changes in the light sensor value as the robot moves over the course illustrated in Figure 8. In this case, the robot crosses the black line on the course several times to measure the characteristics of its light sensor. The sharp dips observed in Figure 7 occur when the robot crosses the black line. While the robot moves over the white part of the course, the sensor value is approximately 900. When it crosses the black line, the sensor value decreases below 400. Students confirm the characteristics of the light sensor by themselves. Based on these results, black and white colors can be identified by using some threshold value, for example, 700.

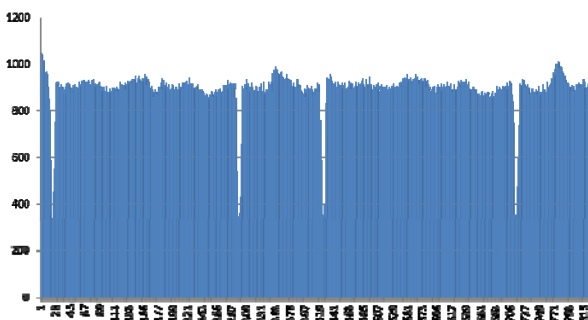


Figure 7: Change of the light sensor value, where horizontal axis is time

If the value of the light sensor is larger than 700, it seems that the robot is running over the white part; otherwise, the robot is running on the black line. Students can know that the robot movement is tracing the black line by using this threshold value.

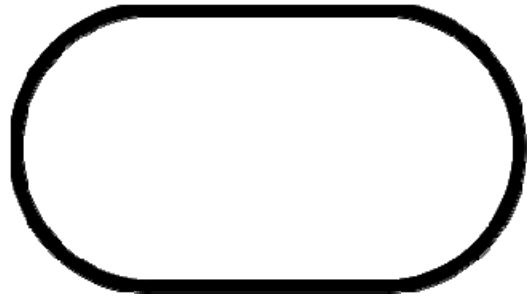


Figure 8: Course for line tracing

5.5 Example 4

Making use of the characteristic of the light sensor and the threshold value, the student can realize line tracing by the robot.

Figure 9 shows the simple program that realizes the line tracing. The variable “target” holds the threshold value. In the infinite “while”, the variable “diff_light” holds the value of the light sensor. If the value of the light sensor is less than the threshold value, the robot makes a right turn; otherwise it makes a left turn. The program repeats this process endlessly every 40 ms. This simple program is able to make the robot move along the black line. Students can learn conditional branching through this example.

```

from nxt_bluetooth import nxt_bt
bt = nxt_bt("00:16:53:0c:48:1e", 0)

print "START"
target = 700
bt.send( "2" )
bt.send( "f" )
while True:
    print bt.diff_light
    if bt.diff_light < target:
        bt.send( "r" )
    else:
        bt.send( "l" )
        bt.wait(0.04)

bt.send( "0" )

```

Figure 9: Example 4—Line tracing by simple control

5.6 The log file

Each time a program is executed, a log file is produced. This file includes a record concerning the details of the robot every 40 ms. Table 3 summarizes the items included in each record of the file. The log data is recorded in CVS format. Figure 10 shows an example of the log file opened in EXCEL. Students can analyze the log file and obtain a

trace of the robot by using some mathematical calculations. Figure 11 shows an example of a trace of the robot obtained from calculations. The trace is almost the same as the course depicted in Figure 8. We can observe zigzag lines in the trace, which is the effect of the simple “ON and OFF” control by the program listed in Figure 9.

Table 3: Items recorded in the log file

Item	Meaning
Time	Elapsed time (ms)
Dt1	PWM value for right motor
Dt2	PWM value for left motor
Batt	Voltage of battery
Mtr1	Rotate angle of tail motor
Mtr2	Rotate angle of right wheel motor
Mtr3	Rotate angle of left wheel motor
ADC s1	Gyro sensor value
ADC s2	Ultrasonic sensor value
ADC s3	Light sensor value
ADC s4	Touch sensor value
I2c	Distance measured by the ultrasonic sensor

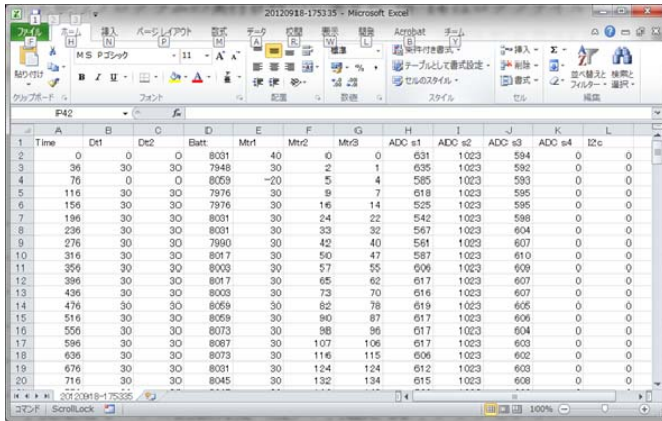


Figure 10: Example of a log file opened in EXCEL

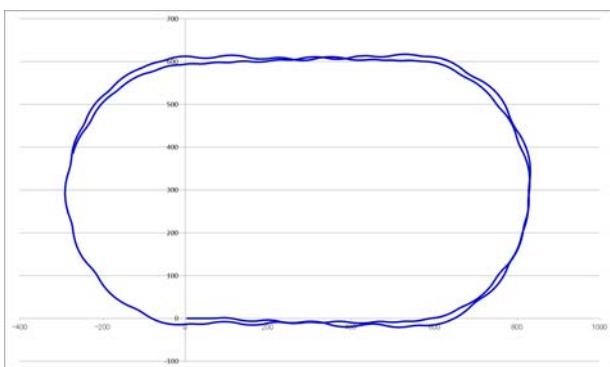


Figure 11: Example of a trace of the robot

6 EXTENSION OF THE PROGRAM FOR CONTROLLING MULTIPLE ROBOTS

The examples explained above concern basic programs that control a single robot. The materials can be applied to the advanced case where multiple robots are controlled by threads. Figure 12 shows the configuration of this case,

where two robots are controlled by a single program. A program developed by a student generates two threads for two robots. Each thread executes the same program and controls one of the two robots. Figure 13 shows the sample program for this.

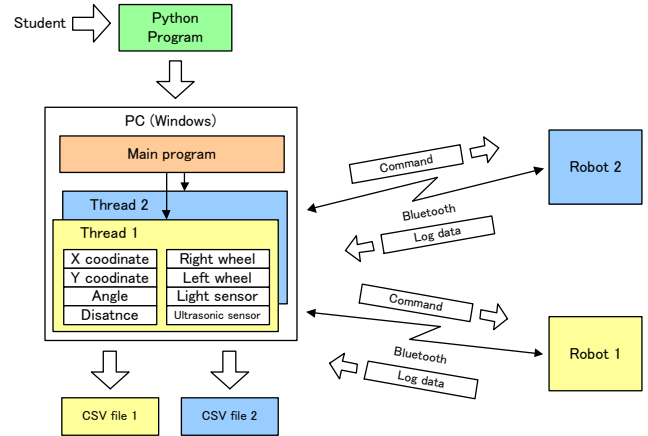


Figure 12: Configuration of the developed materials for controlling two robots

```

import threading # thread model
import time
from nxt_bluetooth import nxt_bt

class test(threading.Thread):
    def __init__(self, s):
        threading.Thread.__init__(self)
        self.setDaemon(True)
        self.bt = nxt_bt(s, 0)

    def run(self):
        self.bt.start()
        self.bt.send( "3" )
        self.bt.send( "f" )
        self.bt.swait(5)
        self.bt.send( "r" )
        self.bt.swait(5)
        self.bt.send( "0" )

if __name__ == "__main__":
    t1 = test("00:16:53:0c:48:1e")
    t2 = test("00:16:53:0c:82:39")
    print "Hit enter key, if you are ready."
    raw_input()
    t1.start()
    t2.start()
    time.sleep(20)
    
```

Figure 13: Program controlling two robots

The part of the program surrounded by the dashed line is the definition of the class that defines the movement of the robots. Two threads are generated from the same class in the main part; the movement of the two robots is the same in this case. Advanced students can learn thread mechanisms through this example.

7 APPLICATION OF THE MATERIALS TO AN ACTUAL CLASS

We applied part of the materials to sessions of an actual experimental class in the first semester of this year. We conducted 9 sessions. The total number of students who participated in the sessions was 45. Approximately half of the students had no experience in programming. The duration of each session was 3 hours and the number of students for one session was at most 7. We explained the material for the first 40 minutes. Two robots were employed to execute the programs. Then students were asked to write four simple programs, including one for line tracing. Although some students had difficulty in understanding Python, 87% of the students indicated a positive impression of the session and expressed their satisfaction when the robot moved correctly. Although students with no programming experience had strong concerns about Python programming, after the session, most of them stated that it was easier than they had thought it would be.

Table 4 shows the summary of their impressions of the materials. Students of Group A had not taken a class in the C programming language as university students, whereas Group B students had, although some students in Group A had learned the C programming language in high school. Students also pointed out aspects of the materials that could be improved.

Table 4: Summary of impressions by students

Impression	Excellent	Good	No comment
Group A	28	1	4
Group B	17	0	1

We also demonstrated the materials and explained a simple program to high school students. A large number of these students found the materials highly interesting.

Through the experience of the actual class, we could identify advantages of the developed materials, summarized as follows:

- The materials could attract more attention from students who had no programming experience.
- As students could edit and execute a Python script directly, program errors were modified quickly. Most of the students could complete the given exercises within the prescribed class hour.
- Students were strongly impressed when the robot performed as they intended.
- Students were also surprised when they got a trace of the robot from the log file.

However, we found drawbacks of the materials from the experience:

- Much effort was needed for preparing and guiding a session.
- Support by teaching assistants was needed for every three or four students to help when they encounter programming problems.
- The difficulty level of programs that students have to develop should be reconsidered. Natural steps from simple to advanced are required.

- We have to improve the assignments before the session to shorten the time needed to explain the materials.
- The number of robots is too small in the case of 7 students.
- Because the quality of components in the robot varied, the robot could not go straight accurately after it received the forward command. Compensation for the error caused this variation in the components is needed.

8 CONCLUDING REMARKS

In this paper we reported the development of teaching materials for computer programming. Our objective is to give beginner students the satisfaction of creating programs that control a robot. Students can implement line tracing by a simple program consisting of a small number of program steps.

Several products target the education of programming by using the robot of LEGO Mindstorms NXT [1]. The typical product is NI LabVIEW for LEGO MINDSTORMS software [8], which makes it possible for students to develop programs by combining predefined blocks graphically. The difference between our materials and this product is that our materials control the robot remotely by the scripting language Python. Students can learn programming through widely used high-level programming language. As a result, they become accustomed to the conventional programming paradigm.

One of the features of the teaching materials is that the module and programs are in a white box state. We are able to flexibly customize the robot itself and the Python module for future requests. Since the characteristics of the students vary depending on the number of students, their interests, their scholastic ability, and characteristics of the university or college, the capability of customizing the materials is considered to be important. We will improve the teaching materials based on the experiences of this semester and extend them for teaching the basics of object-oriented programming.

We are also planning to extend the materials by integrating them with the programming language SCRATCH [7], which is intended for students in elementary or junior high school. The boxes surrounded by the dashed lines in Figure 3 show this extension. We will report on this development in the future.

REFERENCES

- [1] <http://www.etrobo.jp/2013/>.
- [2] <http://www.afrel.co.jp/mindstorms/nxt/>.
- [3] <http://www.python.org/>.
- [4] <http://lejos-osek.sourceforge.net/index.htm>.
- [5] "pybluez," <http://code.google.com/p/pybluez/>.
- [6] "BlueZ," <http://www.bluez.org/>.
- [7] <http://scratch.mit.edu/>.
- [8] <http://www.ni.com/academic/mindstorms/ja/>.

A Design Method of Optimal H_2 Integral Servo Problem

Noriyuki Komine^{*}, Masakatsu Nishigaki^{*}, Kunihiro Yamada^{**} and Tadanori Mizuno^{***}

^{*} Graduate School of Science and Technology, Shizuoka University, Japan
komine@keyaki.cc.u-tokai.ac.jp, nisigaki@inf.shizuoka.ac.jp

^{**} Professional Graduate School of Embedded Technology, Tokai University, Japan
yamada@kunighiroi.com

^{***} Faculty of Information Science, Aichi Institute of Technology, Japan
mizuno@mizulab.net

Abstract -This paper proposes a design method of optimal H_2 integral servo controller. The optimal H_2 integral controller is to establish a way to find the admissible controller such that the controlled plant is stabilized and guarantee to track a constant reference signal while minimizing the H_2 norm of the closed-loop transfer function of the controlled plant from disturbance to the controlled output. The effectiveness of the proposed method controller is verified reducing the torsional vibration of two-inertia system with comparing the traditional optimal servo controller is shown by computer simulation and experimental results.

Keywords: Optimal controller, Integral servo, Torsional vibration

1 INTRODUCTION

In sense of modern control, designing a state feedback for a linear dynamical system which not only stabilizes but also dampens the responses of closed-loop system is generally required [1]. It is also required that the output of a system has no steady-state error for a desired input even if the parameter variations or disturbances exist. Consequently, the integral servo problem was initiated by H. W. Smith and E. J. Davison [2], in which they proposed the state and output integral feedback approaches by the differential transformations, and gave some suggestions on measurement feedback schemes. In addition, optimality in control was primarily concerned by R.E. Kalman [3] to minimize the quadratic performance index of state variables and inputs. These two concepts were then combined and the design method of an optimal tracking system by introducing the integral action for the system using regulator theory was obtained and reported by T. Takeda and T. Kitamori [4]. However, it is difficult to select the proper values of the weighting matrices of performance index in the optimal servo problem to mitigate under damped responses of dynamic systems. Besides, the optimal H_2 servo problem is to find the optimal control such that the output tracks the desired trajectory, minimizing the tracking error cost and state excitation cost in the sense of an optimal H_2 control [5]- [6]. On the other hand, Anderson and Moore [7] introduced an optimal controller with a prescribed degree of stability affecting the locations of all closed-loop poles. However, it dose not

necessarily reduce the under damping of the closed-loop system. Recently, the optimal H_2 control for oscillatory system minimizing a performance criterion involved time derivatives of state vector was formulated to levitate under damped responses of dynamic systems [8]-[11].

In this paper, the optimal H_2 integral servo controller which stabilizes an oscillatory system with the prescribed degree of stability is derived such that the optimal control law is much effective to control an under damped steady-state tracking error by H_2 control framework. The proposed controller obtained from derivative state constrained H_2 integral servo theorem is applied to the two-inertia system. The verification of the effectiveness of the proposed controller in mitigating an under damped responses of dynamic system is also shown in the paper.

2 H_2 INTEGRAL SERVO PROBLEM

In order to obtain the optimal H_2 integral servo controller, the following controlled plant equations are given as

$$\begin{aligned} \frac{d}{dt}x(t) &= Ax(t) + B_2u(t), \quad x(0) = x_0 \\ y(t) &= C_2x(t) \end{aligned} \quad (1)$$

where $x(t)$, $u(t)$ and $y(t)$ denote the state vector, the input vector and the output vector, respectively. The integral $x_I(t)$ of the error vector $e(t)$ between the reference input $r(t)$ and controlled output $y_r(t)$ is defined as

$$x_I(t) = \int_0^t e(\tau) d\tau, \quad e(t) = r(t) - y(t) \quad (2)$$

where $y(t)$ denotes the controlled vector. Using Eq. (1) and Eq. (2), the augmented controlled plant is then given by

$$\begin{aligned} \frac{d}{dt} \begin{bmatrix} x(t) \\ x_I(t) \end{bmatrix} &= \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ x_I(t) \end{bmatrix} + \begin{bmatrix} B_2 \\ 0 \end{bmatrix} u(t) + \begin{bmatrix} 0 \\ r(t) \end{bmatrix} \\ y(t) &= \begin{bmatrix} C_2 & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ x_I(t) \end{bmatrix}. \end{aligned} \quad (3)$$

In order to the steady state tracking error $\lim_{t \rightarrow \infty} e(t)$ should be vanished, the derivative augmented state vector defined as $\frac{dx_I(t)}{dt}$ which should be vanished for approaching infinity of t . The derivative augmented system is given by combining of the derivative state equation of Eq.(1) and the derivative state equation of (3) as

$P(s)$:

$$\begin{cases} \frac{d}{dt} \begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} = \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} + \begin{bmatrix} B_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & B_{1I} & 0 & 0 & 0 & 0 \end{bmatrix} \dot{w}(t) + \begin{bmatrix} B_2 \\ 0 \end{bmatrix} \dot{u}(t) \\ \dot{z}(t) = \begin{bmatrix} C_1 & 0 \\ 0 & C_{1I} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & D_{11I} \\ 0 & D_{11I} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \ddot{x}(t) \\ \ddot{x}_I(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ D_{12} \end{bmatrix} \dot{u}(t) \\ \begin{bmatrix} \dot{y}(t) \\ e(t) \end{bmatrix} = \begin{bmatrix} C_2 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 & D_{21} & 0 \\ 0 & 0 & 0 & 0 & 0 & D_{21I} \end{bmatrix} \dot{w}(t) \end{cases} \quad (4)$$

where $B_1, B_{1I}, C_1, C_{1I}, D_{11}, D_{11I}, D_{12}, D_{21}$ and D_{21I} are denoted the design parameter matrices to obtain the derivative constrained integral servo controller.

The disturbance

$\dot{w}(t) = [\dot{w}_1^T(t) \ \dot{w}_2^T(t) \ \dot{x}^T(t) \ \dot{x}_I^T(t) \ \dot{w}_3^T(t) \ \dot{w}_4^T(t)]^T$ is continuously differentiable in time. By definition of the optimal H_2 integral servo problem, the augmented general plant is given by

$$\begin{cases} \hat{P}_2(s) : \\ \frac{d}{dt} \begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} = \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} + \tilde{B}_1 \dot{w}(t) + \begin{bmatrix} B_2 \\ 0 \end{bmatrix} \dot{u}(t) \\ \dot{z}(t) = \tilde{C}_1 \begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} + \tilde{D}_{12} \dot{u}(t) \\ \begin{bmatrix} \dot{y}(t) \\ e(t) \end{bmatrix} = \begin{bmatrix} C_2 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} + \tilde{D}_{21} \dot{w}(t) \end{cases} \quad (5)$$

where

$$\tilde{B}_1 = \begin{bmatrix} B_1 & 0 & AD_{11} & 0 & 0 & 0 \\ 0 & B_{1I} & -C_2 D_{11I} & 0 & 0 & 0 \end{bmatrix}$$

$$\tilde{D}_{21} = \begin{bmatrix} 0 & 0 & C_2 D_{11} & 0 & D_{21} & 0 \\ 0 & 0 & 0 & D_{11I} & 0 & D_{21I} \end{bmatrix}$$

$$\tilde{C}_1 = \begin{bmatrix} C_1 & 0 \\ 0 & C_{1I} \\ D_{11} A & 0 \\ -D_{1I} C_2 & 0 \\ 0 & 0 \end{bmatrix}, \quad \tilde{D}_{12} = \begin{bmatrix} 0 \\ 0 \\ D_{11} B_2 \\ 0 \\ D_{12} \end{bmatrix}$$

Statement of Derivative State Constrained H_2 integral servo problem:

Let $r(t)$ denote the step reference vector. Derivative State Constrained Optimal H_2 servo integral problem is to find an admissible optimal integral controller such that the controlled plants with augmented integrator is stabilized and the output $y(t)$ tracks the constant reference signal $r(t)$ while minimizing the H_2 norm of the closed-loop transfer function with controlled plant from $L[\dot{w}(t)]$ to $L[\dot{z}(t)]$ of $\hat{P}_2(s)$.

3 SOLUTION OF STATMENT

The solution to the derivative state constrained H_2 optimal control defined above is given by the following procedure.

In order to consider the effect of the prescribed degree of stability α to a controller, each vector variable is exponentially weighted as follows.

$$\begin{bmatrix} \tilde{\dot{x}}(t) \\ \tilde{\dot{x}}_I(t) \end{bmatrix} = e^{\alpha t} \begin{bmatrix} \dot{x}(t) \\ \dot{x}_I(t) \end{bmatrix} \quad (6)$$

$$\begin{cases} \tilde{\dot{w}}(t) = e^{\alpha t} \dot{w}(t) \\ \tilde{\dot{z}}(t) = e^{\alpha t} \dot{z}(t) \\ \tilde{\dot{u}}(t) = e^{\alpha t} \dot{u}(t) \end{cases} \quad (7)$$

$$\begin{bmatrix} \tilde{\dot{y}}(t) \\ \tilde{e}(t) \end{bmatrix} = e^{\alpha t} \begin{bmatrix} \dot{y}(t) \\ e(t) \end{bmatrix} \quad (8)$$

Hence, the generalized plant $\tilde{P}_\alpha(s)$ shown in Eq. (9) after applying the transformed vector variables Eq.(6)-Eq.(8) is given by

$\tilde{P}_\alpha(s)$:

$$\begin{cases} \frac{d}{dt} \begin{bmatrix} \dot{\tilde{x}}(t) \\ \dot{\tilde{x}}_l(t) \end{bmatrix} = \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} + \alpha I \begin{bmatrix} \dot{\tilde{x}}(t) \\ \dot{\tilde{x}}_l(t) \end{bmatrix} + \tilde{B}_1 \dot{\tilde{w}}(t) + \begin{bmatrix} B_2 \\ 0 \end{bmatrix} \dot{\tilde{u}}(t) \\ \dot{\tilde{z}}(t) = \tilde{C}_1 \begin{bmatrix} \dot{\tilde{x}}(t) \\ \dot{\tilde{x}}_l(t) \end{bmatrix} + \tilde{D}_{12} \dot{\tilde{u}}(t) \\ \begin{bmatrix} \dot{\tilde{y}}(t) \\ \dot{\tilde{e}}(t) \end{bmatrix} = \begin{bmatrix} C_2 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \dot{\tilde{x}}(t) \\ \dot{\tilde{x}}_l(t) \end{bmatrix} + \tilde{D}_{21} \dot{\tilde{w}}(t) \end{cases} \quad (9)$$

The solution to the derivative state constrained H_2 optimal control defined above is given by the following procedure.

3.1 Singular value decomposition

There always exist unitary matrices $V_j, U_j, j=1,2$ for the singular value decomposition of \tilde{D}_{12} and \tilde{D}_{21} ;

$$\tilde{D}_{12} = U_1 \begin{bmatrix} 0 \\ \Sigma_1 \end{bmatrix} V_1, \Sigma_1 = \begin{bmatrix} \sigma_{11} & & \\ & \ddots & \\ & & \sigma_{1m_2} \end{bmatrix}, m_2 = \dim(\dot{\tilde{u}}(t))$$

$$\tilde{D}_{21} = U_2 \begin{bmatrix} 0 & \Sigma_2 \end{bmatrix} V_2, \Sigma_2 = \begin{bmatrix} \sigma_{11} & & \\ & \ddots & \\ & & \sigma_{1p_2} \end{bmatrix}, p_2 = \dim \begin{bmatrix} \dot{\tilde{y}}(t) \\ \dot{\tilde{e}}(t) \end{bmatrix}$$

where $\Sigma_i, i=1,2$ are the diagonal singular value matrices. Using the results obtained above, the input and output vectors as well as the generalized plant are accordingly transformed as shown in the following sub-section.

3.2 Variable transformations

The generalized plant can be obtained by using the following variable transformations defined by

$$\dot{\tilde{w}}(t) = V_2 \dot{\hat{w}}(t) \quad (10)$$

$$\dot{\hat{z}}(t) = U_1^T \dot{\tilde{z}}(t) \quad (11)$$

$$\dot{\hat{u}}(t) = V_1 \Sigma_1^{-1} \dot{\tilde{u}}(t) \quad (12)$$

$$\begin{bmatrix} \dot{\hat{y}}(t) \\ \dot{\hat{e}}(t) \end{bmatrix} = \Sigma_2^{-1} U_2^T \begin{bmatrix} \dot{\tilde{y}}(t) \\ \dot{\tilde{e}}(t) \end{bmatrix}. \quad (13)$$

Substituting Eq. (12) and Eq. (13) into Eq. (9), then the transformed generalized plant $\hat{P}_\alpha(s)$ which is reduced to the standard form of the H_2 control problem is then obtained as

$\hat{P}_3(s)$:

$$\begin{cases} \frac{d}{dt} \begin{bmatrix} \dot{\hat{x}}(t) \\ \dot{\hat{x}}_l(t) \end{bmatrix} = \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} \begin{bmatrix} \dot{\hat{x}}(t) \\ \dot{\hat{x}}_l(t) \end{bmatrix} + \hat{B}_1 \dot{\hat{w}}(t) + \hat{B}_2 \dot{\hat{u}}(t) \\ \dot{\hat{z}}(t) = \hat{C}_1 \begin{bmatrix} \dot{\hat{x}}(t) \\ \dot{\hat{x}}_l(t) \end{bmatrix} + \hat{D}_{12} \dot{\hat{u}}(t) \\ \begin{bmatrix} \dot{\hat{y}}(t) \\ \dot{\hat{e}}(t) \end{bmatrix} = \hat{C}_2 \begin{bmatrix} \dot{\hat{x}}(t) \\ \dot{\hat{x}}_l(t) \end{bmatrix} + \hat{D}_{21} \dot{\hat{w}}(t) \end{cases} \quad (14)$$

where

$$\hat{B}_1 = \tilde{B}_1 V_2,$$

$$\hat{B}_2 = \begin{bmatrix} B_2 \\ 0 \end{bmatrix} V_1 \Sigma_1^{-1}$$

$$\hat{C}_1 = U_1^T \tilde{C}_1$$

$$\hat{C}_2 = \Sigma_2^{-1} U_2^T \begin{bmatrix} C_2 & 0 \\ 0 & I \end{bmatrix}$$

$$\hat{D}_{12} = U_1^T \tilde{D}_{12} V_1 \Sigma_1^{-1} = \begin{bmatrix} 0 \\ I \end{bmatrix}$$

$$\hat{D}_{21} = \Sigma_2^{-1} U_2^T \tilde{D}_{21} V_2 = \begin{bmatrix} 0 & I \end{bmatrix}$$

Suppose that the transformed generalized plant $\hat{P}(s)_\alpha$ of Eq. (14) satisfy the following relations:

$$(A1) \left(\begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} + \alpha I, \hat{B}_2, \hat{C}_2 \right)$$

is stabilizable and detectable.

(A2) \hat{D}_{12} and \hat{D}_{21} have full rank.

$$(A3) \left(\begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} + \alpha I, \hat{B}_2, \hat{C}_1, \hat{D}_{12} \right) \text{ has full column}$$

rank for all ω .

$$(A4) \left(\begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} + \alpha I, \hat{B}_1, \hat{C}_2, \hat{D}_{21} \right) \text{ has full row rank}$$

for all ω .

The first assumption (A1) is for the stabilizability of the transformed generalized plant (14). The assumption (A2) is sufficient to ensure the controller is proper. The assumption (A3) and (A4) guarantee two Hamiltonian matrices belong to $\text{dom}(\text{Ric})$.

3.3 Hamiltonian matrices

Under the above assumptions, the optimal H_2 solution to the transformed generalized plant (14) is given as follows;

A couple of Hamiltonian matrices are constituted as

$$H_2 = \begin{bmatrix} \left[\begin{array}{c|c} A & 0 \\ \hline -C_2 & 0 \end{array} \right] + \alpha I + \hat{B}_2 \hat{D}_{12}^T \hat{C}_1 & -\hat{B}_2 \hat{B}_2^T \\ -\hat{C}_1^T \hat{C}_1 + \hat{C}_1^T \hat{D}_{12} \hat{D}_{12}^T \hat{C}_1 & -\left\{ \left[\begin{array}{c|c} A & 0 \\ \hline -C_2 & 0 \end{array} \right] + \alpha I \right\} - \hat{B}_2 \hat{D}_{12} \hat{C}_1^T \end{bmatrix} \quad (15)$$

$$J_2 = \begin{bmatrix} \left[\begin{array}{c|c} A & 0 \\ \hline -C_2 & 0 \end{array} \right] + \alpha I & \left[\begin{array}{c|c} A & 0 \\ \hline -C_2 & 0 \end{array} \right] - \hat{C}_2^T \hat{D}_{12} \hat{B}_1^T & -\hat{C}_2^T \hat{C}_2 \\ -\hat{B}_1 \hat{B}_1^T + \hat{B}_1 \hat{D}_{21} \hat{D}_{21}^T \hat{B}_1^T & -\left\{ \left[\begin{array}{c|c} A & 0 \\ \hline -C_2 & 0 \end{array} \right] + \alpha I \right\} + \hat{C}_2^T \hat{D}_{21} \hat{B}_1^T \end{bmatrix} \quad (16)$$

Then, it is guaranteed that the solutions exist, which make

$$\begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} + \alpha I + \hat{B}_2 \hat{F}_2 \quad \text{and} \quad \begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} + \alpha I + \hat{L}_2 \hat{C}_2 \quad \text{stable.}$$

From the couple of Riccati solutions,

$$X_2 = \text{Ric}(H_2) > 0, \quad Y_2 = \text{Ric}(J_2) > 0 \quad (17)$$

it is able to construct the following optimal solution

$$\hat{K}_{H_2}(s) = \left[\begin{array}{c|c} \left[\begin{array}{c|c} A & 0 \\ \hline -C_2 & 0 \end{array} \right] + \hat{B}_2 \hat{F}_2 + \hat{L}_2 \hat{C}_2 & -\hat{L}_2 \\ \hline \hat{F}_2 & 0 \end{array} \right] \quad (18)$$

to the transformed generalized plant (14), where

$$\begin{aligned} \hat{F}_2 &= -(\hat{B}_2^T X_2 + \hat{D}_{12} \hat{C}_1) \\ \hat{L}_2 &= -(Y_2 \hat{C}_2^T + \hat{B}_1 \hat{D}_{21}) \end{aligned}$$

A general control formulation with the derivative state constrained optimal H_2 integral servo controller $\hat{K}_{H_{2\alpha}}$ is given by the general configuration shown in Figure 1. Consequently, the assumptions supposed above (A1), (A2), (A3) and (A4) can be reduced to the following expressions.

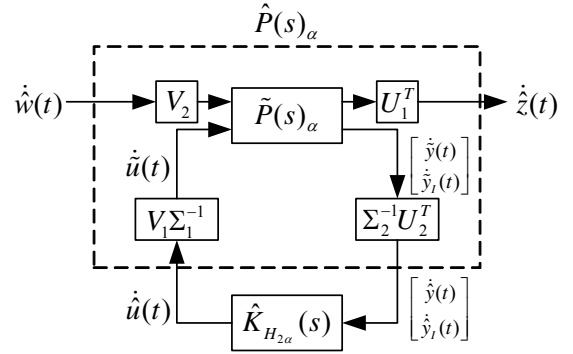


Figure 1: Block diagram of the structure for closed-loop system with equation (18)

Lemma: Suppose the system parameter matrix in equation (14) satisfy the assumptions (A1), (A2), (A3) and (A4), then following assumptions hold;

$$(A1)' \left(\begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix}, \begin{bmatrix} B_2 \\ 0 \end{bmatrix}, \begin{bmatrix} C_2 & 0 \\ 0 & I \end{bmatrix} \right)$$

is stabilizable and detectable.

$$(A2)' \quad D_{12} \quad \text{and} \quad \begin{bmatrix} D_{21} & 0 \\ 0 & D_{21I} \end{bmatrix} \quad \text{have full rank.}$$

$$(A3)' \quad \left(\begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} - j\omega I, \begin{bmatrix} B_2 \\ 0 \end{bmatrix}, \begin{bmatrix} C_1 & 0 \\ 0 & C_{1I} \end{bmatrix}, D_{12} \right) \quad \text{has full column}$$

rank for all ω .

$$(A4)' \quad \left(\begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} - j\omega I, \begin{bmatrix} B_1 & 0 \\ 0 & B_{1I} \end{bmatrix}, \begin{bmatrix} C_2 & 0 \\ 0 & C_I \end{bmatrix}, \begin{bmatrix} D_{21} & 0 \\ 0 & D_{21I} \end{bmatrix} \right) \quad \text{has full}$$

row rank for all ω .

Proof of Lemma: It is clearly shown that the optimal solution for the transformed generalized plant (14) can be obtained under the assumptions (A1)' ~ (A4)', as the following facts of the rank properties, Eq. (19) to Eq. (24).

$$\begin{aligned} & \text{rank} \left[\begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} + \alpha I - j\omega I \quad \hat{B}_2 \right] \\ &= \text{rank} \left[\begin{bmatrix} A & 0 \\ -C_2 & 0 \end{bmatrix} + \alpha I - j\omega I \quad \begin{bmatrix} B_2 \\ 0 \end{bmatrix} \right] \begin{bmatrix} I & 0 \\ 0 & V_1 \Sigma_1^{-1} \end{bmatrix} \quad (19) \\ &= n + r, \quad \forall \text{Re}(s) \geq 0 \end{aligned}$$

$$\begin{aligned}
& \text{rank} \left[\begin{array}{c|c} \left(\begin{array}{cc} A & 0 \\ -C_2 & 0 \end{array} \right) + \alpha I - j\omega I & \\ \hline & \hat{C}_2 \end{array} \right] \\
&= \text{rank} \left[\begin{array}{c|c} \left[\begin{array}{cc} I & 0 \\ 0 & \Sigma_2^{-1} U_2^T \end{array} \right] \left[\begin{array}{cc} A & 0 \\ -C_2 & 0 \end{array} \right) + \alpha I - j\omega I & \\ \hline & \left(\begin{array}{cc} C_2 & 0 \\ 0 & I \end{array} \right) \end{array} \right] \\
&= n + r, \forall \text{Re}(s) \geq 0
\end{aligned} \tag{20}$$

$$\begin{aligned}
& \text{rank } \hat{D}_{12} \\
&= \text{rank } U_1^T \tilde{D}_{12} V_1 \Sigma_1^{-1} \\
&= \text{rank } \tilde{D}_{12} \\
&= m_2
\end{aligned} \tag{21}$$

$$\begin{aligned}
& \text{rank } \hat{D}_{21} \\
&= \text{rank } \Sigma_2^{-1} U_2^T \tilde{D}_{21} V_1 \\
&= \text{rank } \tilde{D}_{21} \\
&= p_2
\end{aligned} \tag{22}$$

$$\begin{aligned}
& \text{rank} \left[\begin{array}{c|c} \left(\begin{array}{cc} A_p & 0 \\ -C_r & 0 \end{array} \right) + \alpha I - j\omega I & \hat{B}_2 \\ \hline & \hat{C}_1 \quad \hat{D}_{12} \end{array} \right] \\
&= \text{rank} \left\{ \left[\begin{array}{cc} I & 0 \\ 0 & U_1^T \end{array} \right] \left[\begin{array}{cc} \left(\begin{array}{cc} A_p & 0 \\ -C_r & 0 \end{array} \right) + \alpha I - j\omega I & \left(\begin{array}{c} B_{2p} \\ 0 \end{array} \right) \\ \left(\begin{array}{cc} C_1 & 0 \\ 0 & C_{1l} \end{array} \right) & (D_{12}) \end{array} \right] \left[\begin{array}{c} I & 0 \\ 0 & V_1 \Sigma_1^{-1} \end{array} \right] \right\} \\
&= \text{rank} \left[\begin{array}{c|c} \left(\begin{array}{cc} A_p & 0 \\ -C_r & 0 \end{array} \right) + \alpha I - j\omega I & \left(\begin{array}{c} B_{2p} \\ 0 \end{array} \right) \\ \hline \left(\begin{array}{cc} C_1 & 0 \\ 0 & C_{1l} \end{array} \right) & (D_{12}) \end{array} \right] \\
&= n + r + m_2, \forall \omega
\end{aligned} \tag{23}$$

$$\begin{aligned}
& \text{rank} \left[\begin{array}{c|c} \left(\begin{array}{cc} A_p & 0 \\ -C_r & 0 \end{array} \right) + \alpha I - j\omega I & \hat{B}_1 \\ \hline & \hat{C}_2 \quad \hat{D}_{21} \end{array} \right] \\
&= \text{rank} \left\{ \left[\begin{array}{cc} I & 0 \\ 0 & \Sigma_2^{-1} V_2^{-1} \end{array} \right] \left[\begin{array}{cc} \left(\begin{array}{cc} A_p & 0 \\ -C_r & 0 \end{array} \right) + \alpha I - j\omega I & \tilde{B}_1 \\ \left(\begin{array}{cc} C_{2p} & 0 \\ 0 & I \end{array} \right) & \tilde{D}_{21} \end{array} \right] \left[\begin{array}{c} I & 0 \\ 0 & V_2 \end{array} \right] \right\} \\
&= \text{rank} \left[\begin{array}{c|c} \left(\begin{array}{cc} A_p & 0 \\ -C_r & 0 \end{array} \right) + \alpha I - j\omega I & \tilde{B}_1 \\ \hline \left(\begin{array}{cc} C_{2p} & 0 \\ 0 & I \end{array} \right) & \tilde{D}_{21} \end{array} \right] \\
&= n + r + p_2, \forall \omega
\end{aligned} \tag{24}$$

This concludes the proof of the Lemma.

4 MAIN RESULT

By integrating the transformed generalized plant (14) with respect to time t with all initial values equal to zero, the optimal servo controller is obtained by following theorem. Thus, the optimal H_2 servo control solution for the system (14) is given by Eq.(18) of the theorem under the assumptions (A1)', (A2)', (A3)' and (A4)'. We have the following main result.

Theorem (Derivative State Constrained Optimal H_2 Integral Servo.)

The derivative state constrained H_2 integral servo controller for the controlled plant (5) is given as

$$K_{H_{2\alpha}}(s) = \left[\begin{array}{c|c} \left[\begin{array}{cc} A & 0 \\ -C_2 & 0 \end{array} \right] + \left[\begin{array}{c} B_2 \\ 0 \end{array} \right] F_{2\alpha} + L_{2\alpha} \left[\begin{array}{cc} C_2 & 0 \\ 0 & I \end{array} \right]_2 & -L_{2\alpha} \\ \hline F_{2\alpha} & 0 \end{array} \right] \tag{25}$$

under the assumptions (A1)', (A2)', (A3) and (A4)', where

$$\begin{aligned}
F_{2\alpha} &= V_1 \Sigma_1^{-1} F_2 = -V_1 \Sigma_1^{-1} V_1^T \left\{ \left[\begin{array}{cc} B_2^T & 0 \end{array} \right] X_2 + D_{12}^T \tilde{C}_1 \right\} \\
L_{2\alpha} &= L_2 \Sigma_2^{-1} U_2^T = - \left\{ Y_2 \left[\begin{array}{cc} C_2^T & 0 \\ 0 & I \end{array} \right] + \tilde{B}_1 \tilde{D}_{21}^T \right\} U_2 \Sigma_2^{-1} U_2^T
\end{aligned}$$

This theorem can be proved as follow.

Proof: As the facts of the rank properties of Eq.(19)-Eq.(24), this immediately shows that the optimal solution (25) for the generalized plant (5) implies the theorem under the assumptions of (A1)', (A2)', (A3)' and (A4)'. This concludes the proof of the theorem. \square

5 SIMULATION AND EXPERIMENTAL RESULTS

A torsional vibration is occurred to the speed of motor by connecting flexible shaft. The vibration is an impediment to improve the characteristics of the two-inertia system. The simulation and experimental results of the speed control of the two-inertia system using the proposed controller will be shown in this section. A structure of two-inertia system is shown in Figure 2.

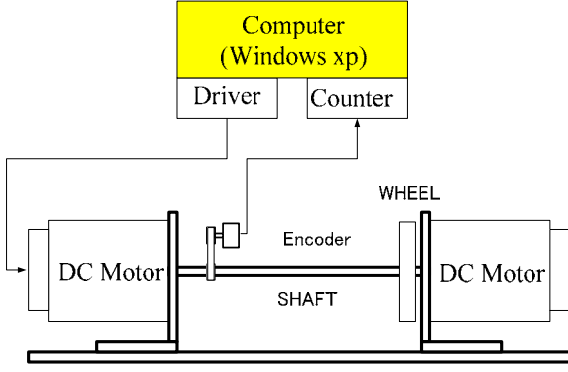


Figure 2: Structure of the two-inertia system

The linear dynamic equation of the two-inertia resonant system with constant disturbance T_L is represented by

$$\begin{aligned} J_m \frac{d\omega_m}{dt} + F_m \omega_m &= u(t) + \tau_d \\ J_L \frac{d\omega_L}{dt} + F_L \omega_L &= \tau_d - T_L \end{aligned} \quad (26)$$

$$\frac{d\tau_d}{dt} = K_s (\omega_m - \omega_L)$$

where J_m, J_L, F_m, F_L and K_s are the inertia of motor, the inertia of load, the friction of motor, friction of load and spring constant of the shaft, respectively. The integral $x_I(t)$ of the error vector $e(t)$ between the reference input $r(t)$ and controlled output $\omega_m(t)$ is defined as

$$\frac{d}{dt} x_I(t) = e(t) = r(t) - y(t) \quad (27)$$

The augmented controlled plant (3) is then given by

$$\frac{d}{dt} x(t) = \begin{bmatrix} F_m/J_m & 0 & 1/J_m & 0 \\ 0 & -F_L/J_L & 1/J_L & 0 \\ K_s & -K_s & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 1/J_m \\ 0 \\ 0 \\ 0 \end{bmatrix} u(t) + \begin{bmatrix} 0 \\ 0 \\ 0 \\ r(t) \end{bmatrix} \quad (28)$$

$$y(t) = [1 \ 0 \ 0 \ 0] x(t)$$

where, $x(t) = [\omega_m(t) \ \omega_L(t) \ \tau_d(t) \ x_I(t)]^T$, $\omega_m(t)$ denotes the speed of motor at time t , $\omega_L(t)$ denotes the speed of load at time t and $\tau_d(t)$ represents the torque of shaft. The numerical values of J_m, J_L, K_s are shown in Table 1. In the case of the numerical values, the friction of motor, friction of load and spring constant of the shaft are neglected, respectively.

Table 1: Numerical values of two-inertia system

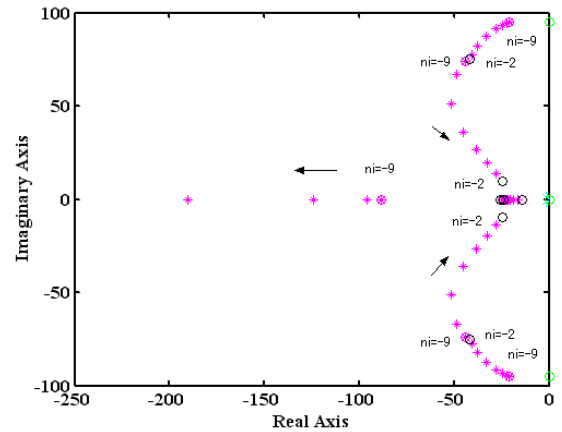
$J_m [Kg \cdot m^2]$	$J_L [Kg \cdot m^2]$	$K_s [N/m]$
0.0866	0.0866	400

The designing parameters

$B_1, B_{1I}, C_1, C_{1I}, D_{11}, D_{11I}, D_{12}, D_{21}$ and D_{21I} in the generalized plant of Eq.(4) are chosen as:

$$\left. \begin{aligned} C_1 = B_1^T &= \text{diag}[\sqrt{10^{qi}} \ \sqrt{10^{qi}} \ \sqrt{10^{qi}}] \\ C_{1I} = B_{1I}^T &= [20000] \\ D_{11} &= \text{diag}[\sqrt{e^{ni}} \ \sqrt{e^{ni}} \ \sqrt{e^{ni}}] \\ D_{11I} &= [100] \\ D_{12} &= [1] \\ \left[\begin{array}{c} D_{21} \\ D_{21I} \end{array} \right] &= \text{diag}[\sqrt{0.01} \ \sqrt{0.01}] \end{aligned} \right\} \quad (29).$$

The variation of closed-loop poles when ni varying from $ni = -9$ to $ni = -2$ is shown in Figure 3. It is seen that the original poles of the open-loop system locate on the imaginary axis. It verifies that the pair of poles with imaginary part approach to the real axis when the parameter ni becomes large.


 Figure 3: Closed-loop poles location for ni varying from $ni=-9$ to $ni=-2$

The simulation results for step responses of the speed of motor with step disturbance shown in Figure 4 clearly explain the effectiveness of the proposed controller when the reference speed of the two-inertia system and the prescribed degree of stability α are assigned to be 1500 rpm and 20, respectively. Significantly, the torsional resonance of two-inertia system is removed when the designing parameter ni is equal to -2 as shown by the dotted line. It is also seen that the torsional resonance of two-inertia system cannot be rejected by $ni = -9$.

On the other hand, the effectiveness of the prescribed degree of stability is also verified by simulation when the designing parameter ni is equal to -2. The speed responses of two-inertia system at the reference speed 1500 rpm are shown in Figure 5. In the case of $\alpha = 20$, it is seen that the

speed of motor can reach the target speed at 1500 rpm rapidly than $\alpha = 0$.

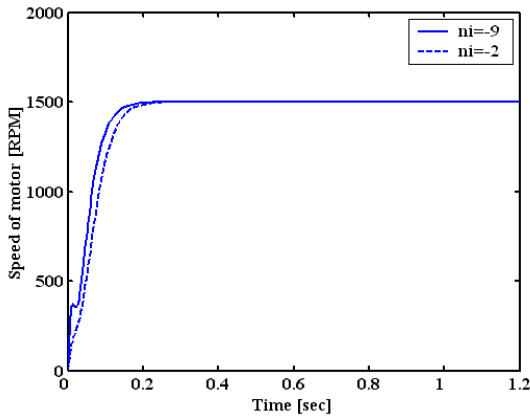


Figure 4: Step responses for $ni = -2$ and $ni = -9$ when $\alpha = 20$

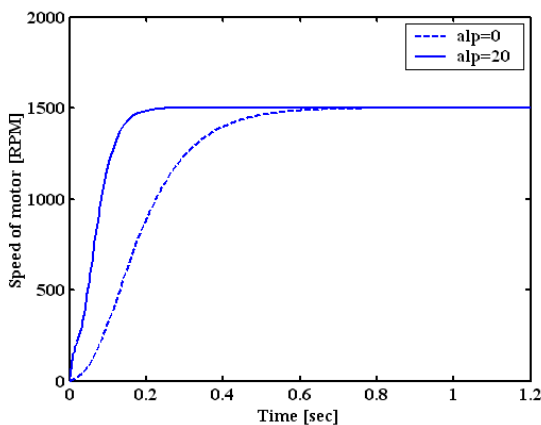


Figure 5: Step responses for $\alpha = 0$ and $\alpha = 20$ when $ni = -2$

The proposed controller is implemented to control the speed of the motor of the two-inertia system with the same condition used in simulation. It is shown that the effectiveness of the controller can be confirmed by the experimental results shown in Figure 7. In Figure 6, the oscillatory response occurred for selecting the weak design parameters $ni = -10$ in Eq.(29) as

$$D_{11} = \text{diag}[\sqrt{e^{-10}} \quad \sqrt{e^{-10}} \quad \sqrt{e^{-10}}].$$

However, in Figure 7, the oscillatory response can be reduced for selecting the design parameter $ni = 0$ as

$$D_{11} = \text{diag}[\sqrt{e^0} \quad \sqrt{e^0} \quad \sqrt{e^0}].$$

Significantly, the torsional resonance of two-inertia system is removed by the designing parameter ni as shown in Figure 7.

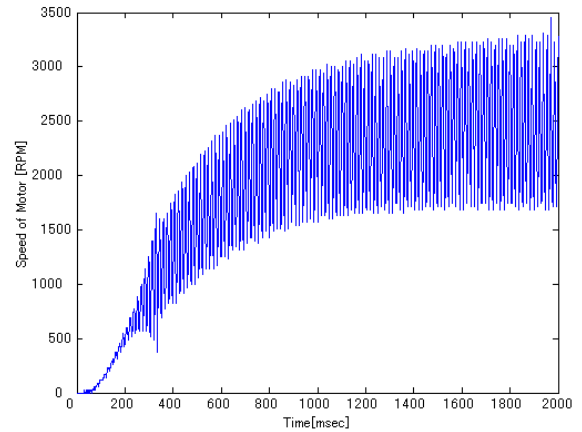


Figure 6: Response of speed of motor for $ni = -10, qi = 0$ and $\alpha = 0$

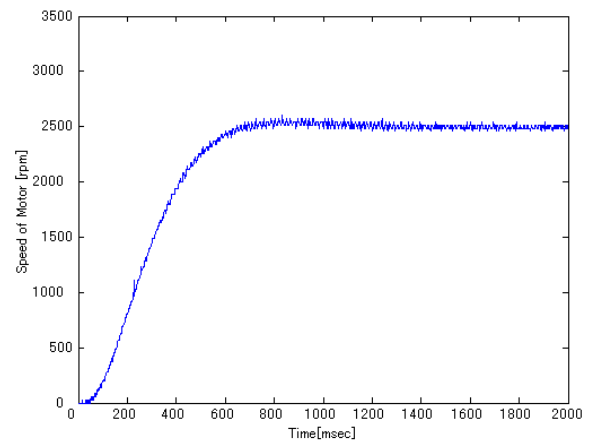


Figure 7: Response of speed of motor for $ni = 0, qi = 2$ and $\alpha = 0$

6 CONCLUSION

The optimal H_2 integral controller using derivative state constrains has been proposed. The proposed controller is effective to control an under damped responses of the controlled system by H_2 control framework. It is recognized that the controller can be applied to the systems whose reference inputs as well as disturbances are all given by step functions. The simulation and experimental results have verified that the proposed schemes can be applied to reduce the oscillation for the two-inertia system.

REFERENCES

- [1] J. J. Hench, C. He, V. Kucera and Meheman, "Damping Controllers via Riccati Equation Approach," IEEE Tran. On AC, Vol.43 No.9, 1280/1284 (1998)
- [2] H. W. Smith and E. J. Davison, "Design of Industrial Regulators," Proc. IEE, Vol. 119, No. 8, pp. 1210/

- 1216, (1972)
- [3] R. E. Kalman, "When Is a Linear Control System Optimal?" ASMEJ. Basic Engineering, ser. D, Vol. 86, 51/60, (1964)
 - [4] T. Takeda and T. Kitamori, "A Design Method of Linear Multi-Input-Output Optimal Tracking Systems," Trans. SICE, Vol. 14, No. 4, 359/364, (1978)
 - [5] M. Ikeda and N. Suda, "Synthesis of Optimal Servosystems," Trans. SICE, Vol. 24, No.1, 40/46, (1988)
 - [6] I. Masubuchi, A. Ohara and N. Suda, "A Design of Robust Servo Systems for Structured Uncertain Plants," Proc. SICE, Vol. 30, No. 9, 1051/1059, (1994)
 - [7] B. D. O Anderson and J. B. Moore, "Linear System Optimization with Prescribed Degree of Stability," Proc. IEE, Vol. 116, No. 12, 2083/2089, (1969)
 - [8] T. Trisuwannawat, K. Tirasesth, M. Iida, N. Komine and Y. Ochiai, "Derivative State Constrained Optimal H_2 Control for Oscillatory Systems and Its Application," Trans. IEEJ (Industry Applications Society; IAS), Vol. 120-D, No. 6, 775/781, (2000)
 - [9] N. Komine and K. Yamada, "Optimal H_2 Integral Controller Design with Derivative State Constraints for Torsional Vibration Model," KES2010, Springer-Verlag, Cardiff, Wales, UK. (2010)
 - [10] W. Ichiyama, N. Komine, T. Benjanarasuth and M. Yoshida, "Optimal H_2 Integral Servo Controller with Derivative State Constraints for Suppressing under Damping of Oscillatory System" JSST2011, Tokyo Japan (2011)
 - [11] N. Komine, M. Nishigaki, T. Mizuno and K. Yamada, "A Design Method of Derivative State Constrained H_2 Integral Servo Controller for Suppressing under Damping of Oscillatory System" AsiaSim2012, Part I, Springer-Verlag Berlin Heidelberg (2012)

Subgrid Search Algorithm for Solving Hitorinishitekure

Shohei Okuyama* and Naoya Chujo*

*Department of Information Science, Aichi Institute of Technology, Japan
{b13711bb, ny-chujo}@aitech.ac.jp

Abstract - Hitorinishitekure, meaning "leave me alone" in Japanese, is a simple puzzle played on an $n \times n$ grid. All cells of the grid contain numbers. The goal of Hitorinishitekure is to paint some cells black in such a way that no number is repeated in any column or row. There are two rules to follow in painting cells: No two black cells should share an edge, and black cells should not divide the $n \times n$ grid into separate parts.

In this paper, the subgrid search algorithm is proposed as a method for solving Hitorinishitekure quickly. Three kinds of subgrid, I-type, L-type, and S-type, are used to reduce the search space.

The algorithm is evaluated using forty 8×8 and forty 12×12 puzzles. On average, Subgrid Search solves 8×8 and 12×12 puzzles in 0.014 and 0.099 seconds, respectively. We developed a smartphone application which solves printed puzzles by using number recognition.

1 INTRODUCTION

Hitorinishitekure is a puzzle game played on an $n \times n$ grid [1]. In this puzzle, all cells of the grid contain numbers and the goal is to paint some cells black in such a way that no number is repeated in the same column or row. There are two rules in painting cells: No black cells share an edge, and black cells should not divide the grid into separate parts.

Hitorinishitekure is similar to Sudoku. Sudoku is expressed as Constraint Satisfaction Problem (CSP) and can be solved by Boolean satisfiability (SAT) solvers [2][3]. The performance improvements of SAT solvers in the last decade are remarkable, and SAT has been a widely used for solving practical applications [4][5][6]. Sugar (SAT-based constraint solver) [7] shows excellent performance for solving various kinds of puzzles including Hitorinishitekure.

In this paper, we propose a heuristic algorithm for solving Hitorinishitekure. It has been implemented in a smartphone application that handles printed puzzles. Hitorinishitekure and its rules are introduced using examples in Section 2. The proposed algorithm is described in Section 3. The performance of the algorithm is evaluated in Section 4. The smartphone application using number recognition is introduced. This work is discussed in section 5 summarized in Section 6.

2 RULES

Hitorinishitekure has the following three rules for painting cells black.

1. No number should repeat in the same column or row.
2. No black cells share an edge.

3. Black cells should not divide the $n \times n$ grid into separate parts.

There is no restriction in the grid size. Commonly used sizes include 4×4 , 8×8 , and 12×12 . Since all repeated cells are considered for being painted black, an upper bound on the number of patterns searched is 2^m where m denotes the numbers of repeated cells in the grid. Thus, the puzzle difficulty increases as the grid size increase.

Figure 1 shows an example of the puzzle. The example consists of a 4×4 grid with all cells of the grid containing numbers. Some numbers repeat in some rows and columns. For example, both cell (1 1) and cell (1 3) contain the same number, one, in the first row, and both cell (2 1) and cell (3 1) contain the same number, two, in the first column. The goal is to paint such repeating cells black in such a way that no number is repeated in any column or row. Figure 1(b) shows the solution. Cell (1 1), cell (2 4), cell (3 1), and cell (4 3) are painted black.

1	3	1	2		3	1	2
2	1	3	4	2	1	3	
2	4	2	1		4	2	1
3	2	2	4	3	2		4

(a) puzzle (b) solution
Fig. 1: Example of Hitorinishitekure

No repeating cells are in the solution (rule 1), and no black cells share an edge (rule 2).

To check rule 3, spanning trees are constructed by connecting remaining white cells as vertices. When two or more trees exist, the grid is defined to be divided. It is clear that one spanning tree can be constructed by remaining white cells. Black cells do not divide the $n \times n$ grid into separate parts (rule 3).

3 PROPOSED ALGORITHM

This section describes the proposed method for solving the puzzle.

3.1 Subgrid Search

In the proposed method, a subgrid is a series of cells that have specific number patterns. Three types of Subgrid are defined:

- I-type: linearly oriented cells containing the same number in the start and end cells. (1) Two cells of the same number sharing an edge. (2) Two or more cells of the same number sharing no edge

- L-type: L-shaped series of cells containing the same number in the start, end, and corner cells. The other cells beside the start, end, and corner cells containing any numbers but the number of the start, end, and corner cells
- S-type: four cells arranged in a square containing the same number

In addition, black cells and white cells are defined:

- Black cell: cell determined to be painted black
- Black cells have to be checked by rule 2.
- White cell: cell undetermined to be painted black.
- White cells have to be checked by rule 1 and 3.

Since the painting patterns of a subgrid are limited by the rules, the search space can be reduced. The three types of subgrid are explained in order.

Figure 2 shows examples of the first type of subgrid, I-type. Cells from (2 2) to (2 6) form an I-type subgrid. Cells (2 3) and (2 5) can contain any number but one. Only one of cells (2 2), (2 4), and (2 6) should be a white cell; the remaining two cells should be black cells. The number of patterns to be searched is three. The number of patterns to be searched by the brute-force method is eight (2^3). Thus, the search space size is reduced by a factor of $3/8$ by identifying this I-type Subgrid.

When an I-type subgrid consists of n cells having the same number, the number of patterns to be searched is n , whereas the number is 2^n by the brute-force method. Thus, the search space size is reduced by a factor of $n/(2^n)$ by identifying this I-type subgrid.

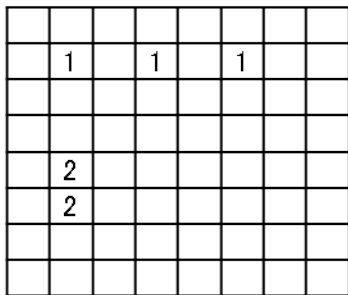


Fig. 2: Examples of I-type subgrids

Figure 3 shows examples of the second type of subgrid, L-type. Cells (2 2), (2 4), and (4 2) define an L-type subgrid, which has three cells with the same number.

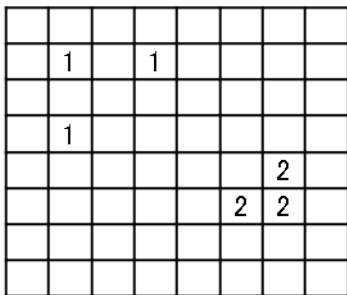


Fig. 3: Examples of L-type subgrids

Cells (2 3) and (3 2) can contain any number but one by the definition. There are two cases of cell painting. When cell (2 2) is black (white), the (2 4) and (4 2) cells must be white (black). In this example, the number of patterns to be searched is two. The number to be searched by the brute-force method is 2^3 . Thus, the search space size is reduced by a factor of $1/4$ by identifying this L-type subgrid.

Two patterns are used in the proposed algorithm in order to satisfy rule 1. However, remaining white cells, like (2 4) and (4 2) in the example may be painted black in the later search. It depends on other cells' placements.

Figure 4 shows examples of the third type of subgrid, S-type. There are two cases of cell painting. In the left-hand example, when cells (2 2) and (3 3) are black (white), cells (2 3) and (3 2) are white (black). The number of patterns to be searched by the brute-force method is 2^3 . Thus, the search space size is reduced by a factor of $1/4$ by identifying this S-type subgrid.

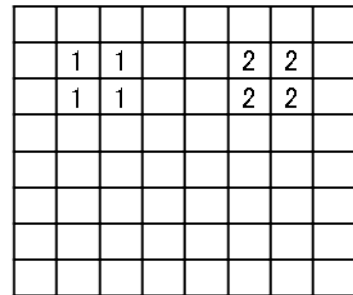


Fig. 4: Examples of S-type subgrids

Repeating cells always belong to an I-type subgrid, but this subgrid can change to L-Type or S-type by adding other cells.

3.2 Algorithm

In this subsection, the algorithm for solving the puzzle is explained using examples. It consists of the following steps.

1. Processing of self-determined cells
2. Reduction by Subgrid search
3. Brute-force search

Firstly, this algorithm searches for self-determined cells. Then identified self-determined cells are determined as a black or white cell from the initial number placement. Secondly, the algorithm searches for any Subgrid which is adjacent to a black cell. For any subgrid found, the adjacent cell in the subgrid is determined to be a white cell by rule 2. The other cells in the Subgrid are determined to be white or black. The algorithm keeps searching until it finds all subgrids adjacent to a black cell.

Finally, other cells, including those in remaining subgrids, are processed to find a solution by a brute-force search.

3.2.1. Process of Self-Determined Cells

The initial puzzle includes self-determined cells. Self-determined cells are defined to be white or black cells without search by rules.

Figure 5 shows examples of self-determined cells. The three cells (2 2), (2 3), and (2 4) have the same number, one. Both the end cells, (2 2) and (2 4), must be the cells to be painted black because of rule 2.

The three cells (4 1), (4 5), and (4 6) have the same number, three. In this case, cell (4 1) is determined to be a black cell because either cell (4 5) or (4 6) should be a white cell by rule 2. In the same manner, cell (6 7) is determined to be a white cell by rule 2, and so cell (6 4) is determined to be a black cell.

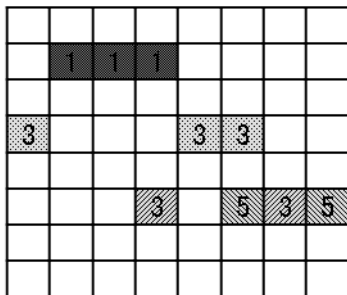


Fig. 5: Examples of self-determined cells

3.2.2. Reduction by Subgrid search

Reduction by subgrid search is explained using an example. Let Figure 6 show the result of the processing of self-determined cells. Blank cells have had numbers omitted for simplicity.

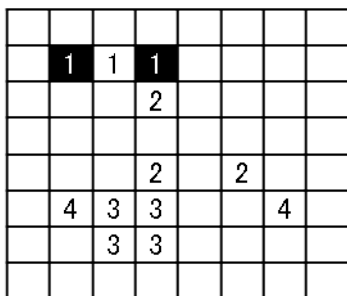


Fig. 6: First stage of reduction by subgrid search

In this process, subgrids adjacent to black cell (2 4) are searched for. An L-type Subgrid defined by cells (3 4), (5 4), and (5 7) is found. Cells (3 4) and (5 7) are determined to be white cells, and cell (5, 4) is determined to be a black cell. Figure 8 shows the results.

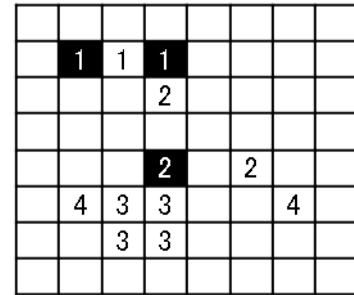


Fig. 7: Second stage of reduction by subgrid search

In the next stage, subgrids adjacent to black cell (5 4) are searched for. An S-type subgrid defined by cells (6 3), (6 4), (7 3), and (7 4) is found. Cells (6 4) and (7 3) are determined to be black cells, and cells (6 3) and (7 4) are determined to be white cells. Figure 9 shows the results.

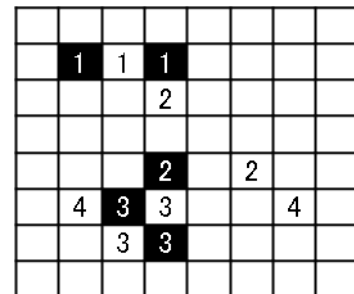


Fig. 8: Third stage of Reduction by subgrid search

In the next stage, subgrids adjacent to black cell (6 3) are searched for. An I-type subgrid defined by cells (6 2) and (6 7) is found. Cell (6 2) is determined to be a white cell, and cell (6 7) is determined to be a black cell. Figure 9 shows the results.

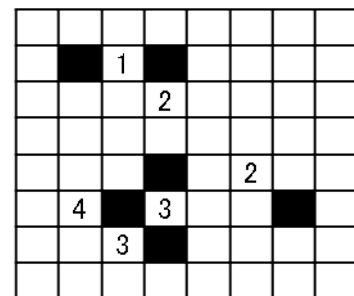


Fig. 9: Fourth stage of reduction by subgrid search

3.2.3. Brute-force search

Cells remaining after subgrid search are processed by brute-force search [4]. Let Figure 10 shows a state of before brute-force search.

1	7	2		3	8		4
6		1	3		2	5	8
	1		6	7	5		3
5	3	4		8		1	6
3		8	1	5	4	6	
	4		5		6		7
4	8	5	2	6		7	1
2		3		1	7	4	8

Fig. 10: State of before brute-force search

There are two cells, (2 8) and (8 8), having the same number, eight, in the eighth column. Either cell can be painted black. However, if cell (2 8) were to be painted black, the $n \times n$ grid would be divided into two parts, which would be in violation of rule 3. Thus, brute-force search finds that the other cell, (8 8), should be painted black. Figure 11 shows the result.

1	7	2		3	8		4
6		1	3		2	5	8
	1		6	7	5		3
5	3	4		8		1	6
3		8	1	5	4	6	
	4		5		6		7
4	8	5	2	6		7	1
2		3		1	7	4	

Fig. 11: Found solution

4 EXPERIMENTS

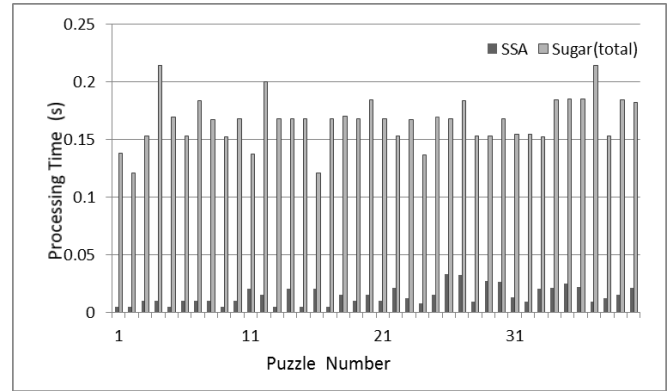
In this section, we present experimental results to demonstrate how well the proposed method works. We implement the proposed method using Java (Ver. 7) and evaluate it on a Microsoft Windows 7 computer with an Intel Core 2 duo CPU (2.40 GHz, 2 GB RAM).

The proposed method was evaluated using examples, which consist of forty puzzles each of 8x8 and 12x12 grids. On average, the algorithm required 0.014 and 0.099 seconds to solve 8x8 and 12x12 puzzles, respectively.

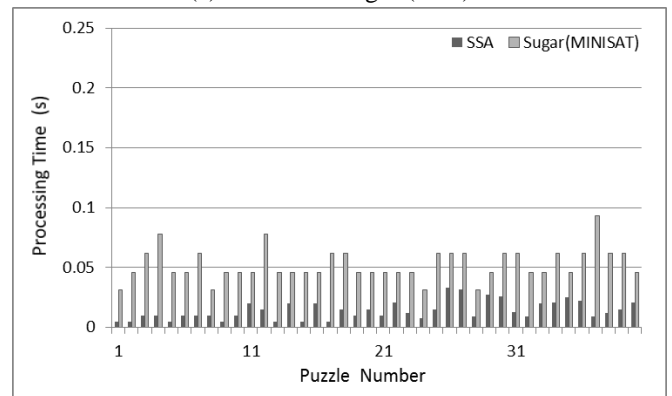
The results are compared with the results by Sugar (SAT-based constraint solver) [7] using the same forty puzzles. On average, Sugar required 0.166 and 0.408 seconds to solve 8x8 and 12x12 puzzles, respectively. This total time includes encoding CSP to SAT, solving SAT and decoding SAT output. The unit time of solving SAT required 0.052 and 0.280 seconds on average, respectively.

Figure 12 shows the comparison results for solving 8x8 puzzles, where SSA denotes Subgrid Search Algorithm. In Figure 12(a), the total time is shown. In Figure 12(b), the unit time by MINISAT (SAT solver) is shown. Figure 13 shows the comparison results for solving 12x12 puzzles.

The proposed algorithm was applied to solve 17x17 puzzles by the proposed algorithm, but most of them could not be solved.

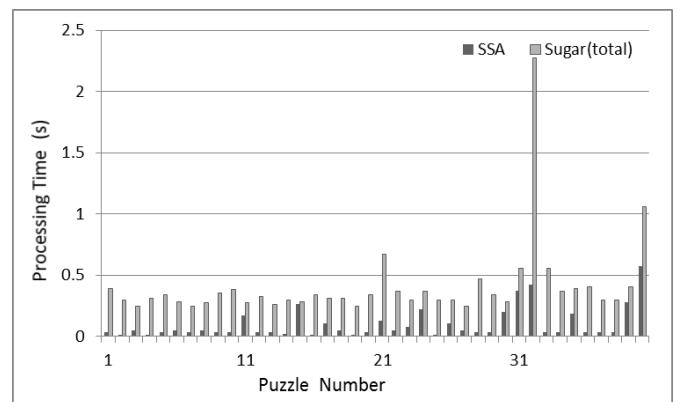


(a) SSA and Sugar (total)

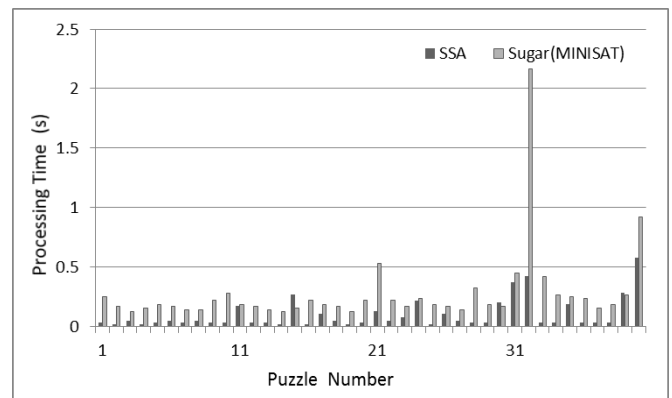


(b) SSA and Sugar (MINISAT)

Fig. 12: Comparison of processing time of 8x8 puzzles



(a) SSA and Sugar(total)



(a) SSA and Sugar(total)

(b) SSA and Sugar(MINISAT)

Fig. 13: Comparison of processing time of 12x12 puzzles

We have developed a smartphone application which solves printed puzzles by number recognition. Figure 14 shows the flow to obtain problems from a captured image.

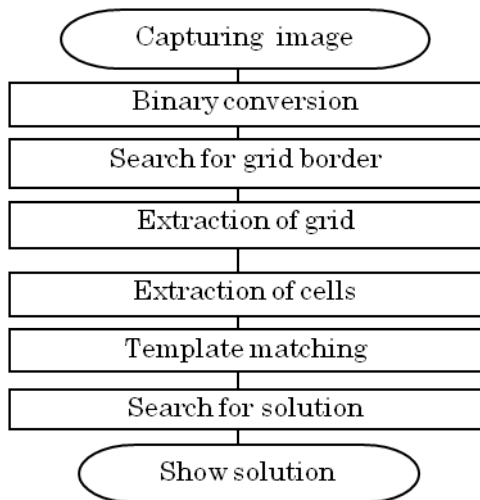


Fig.14: flow of image recognition solution

Figure 15 shows an example of a captured image and the output image by the developed smartphone application. Zeros in the output image denote black cells.

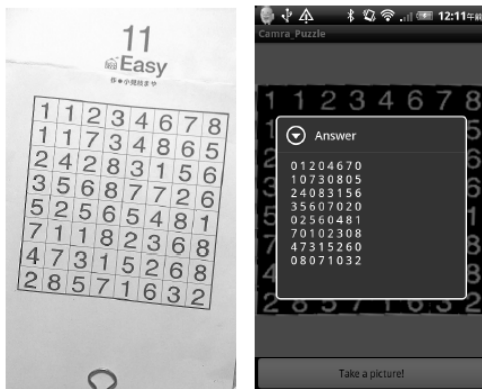


Fig.15: Captured image and output image

5 CONSIDERATION

The proposed method for solving Hitorinishitekure is a modification of the brute-force method. Since search space size for the brute-force method grows exponentially, it is important for us to know how to reduce the search space before applying brute-force search.

The proposed method reduces the search space by introducing subgrids. Experimental results show that the proposed method using three types of subgrid (I-type, S-type, and L-type) is effective.

The proposed method solves Hitorinishitekure faster than Sugar (SAT-based Constraint Solver) for 8x8 and 12x12 puzzles.

However, Sugar solves 17x17 puzzles [7], which could not be solved by the proposed algorithm. This means that the

proposed heuristics are not enough for 17x17 or larger puzzles.

6 CONCLUSIONS

In this paper, we proposed a fast method for solving Hitorinishitekure. We modified the brute-force method in order to reduce the search space by introducing subgrid search. Three types of subgrid (I-type, L-type, and S-type) are implemented.

The experimental results demonstrate that Subgrid Search is effective for solving 8x8 and 12x12 Hitorinishitekure.

We have developed an Android application based on the proposed method which solves printed problems by using image recognition.

REFERENCES

- [1] Nikoli; Nikoli's Official Puzzle Guide of Hitorinishitekure, <http://www.nikoli.co.jp/ja/puzzles/hitori.html> (6/15/2013)
- [2] Lynce, I., Ouaknine, J.: Sudoku as a SAT problem, Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics, 2006.
- [3] Weber, T.: A SAT-based Sudoku solver, Proceedings of the 12th International Conference on Logic for Programming, Artificial Intelligence and Reasoning, pp.11-15, 2005.
- [4] Marques-Silva, J.: Practical Applications of Boolean Satisfiability, Workshop on Discrete Event Systems (WODES), 2008.
- [5] Inoue, K. and Tamura, N.: Editors' Introduction of Recent Advances in SAT Techniques, Journal of Japanese Society for Artificial Intelligence, Vol. 25, No. 1, pp. 56-57, 2010.
- [6] Umemura, A.: SAT/SMT solvers and their Application, Computer Software, Vol. 27, No. 3, pp. 24-35, 2010.
- [7] Tamura, N. and Banbara M.: Sugar: A CSP to SAT Translator Based on Order Encoding, Proceedings of the Second International CSP Solver Competition, pp. 65-69, 2008. <http://bach.istc.kobe-u.ac.jp/sugar> (6/15/2013)

Verification of a Control Program for a Line Tracing Robot using UPPAAL Considering General Aspects

Toshifusa Sekizawa[†], Kozo Okano[‡], Ayako Ogawa[†], and Shinji Kusumoto[‡]

[†]Faculty of Informatics, Osaka Gakuin University, Japan

[‡]Graduate School of Information Science and Technology, Osaka University, Japan

Abstract -

The demand for embedded systems have increased in our society. Ensuring the safety properties of these systems has also become important. Model checking is a technique to ensure such systems. Our target is formal verification of hybrid systems which contain both continuous and discrete behaviors. For the goal, we have studied properties of a line tracing robot built using LEGO Mindstorms with a control program written in LeJOS. We have already presented verification of safety properties of a control program for the application using model checker UPPAAL. In the previous study, we were in a preliminary stage and set limitations. In this presentation, we extend our previous study. In general, a real course can be expressed in combinations of straight and arc courses. First, we verify properties of the same control program for arc courses. Next, in case of the line tracer can not keep track, we analyze turning angle using counter examples. Above-mentioned two approaches are necessary from the standpoint of design phase.

Keywords: Embedded Systems, Formal Verification, Timed Automaton

1 INTRODUCTION

The demand for embedded systems have increased in our society. In these circumstances, it is important to ensure safety properties of embedded systems. Formal methods are mathematical based techniques for verification and development. Model checking is one of formal methods and is widely used in order to ensure properties. Model checking techniques take model and logical formula as their input. Given a model that represents a system under consideration, model checking automatically determines whether or not the model satisfies a given property by exhaustively searching for the state space of the model.

There are various kinds of model checking techniques. Most model checking techniques are based on the finite state machine. For example, a conventional model checking is based on Kripke structure and only deals with discrete variables. However, some embedded systems require time properties in their specification. Several models have been proposed to deal with such real-time systems. One of such approaches is the timed automaton [1]. Timed automaton uses clock variables which range over real numbers. Therefore, timed automaton model can naturally represent the behavior of real-time systems. One of major verifier for timed automaton is UPPAAL [2] in which extended timed automata is used to construct models. UPPAAL can deal with bounded integer vari-

ables and guard expressions on transitions which allow expressions of constraints on variables.

Embedded systems sometimes consist of continuous and discrete dynamics. Such systems are called hybrid systems [3]. We are motivated to verify the behavior of embedded control systems. Especially if these embedded systems are considered to be hybrid systems. For the goal, we have presented verification of safety property of a control program for a line tracing robot using model checker UPPAAL. In the previous study, we were in a preliminary stage and set limitations. For example, we only considered a straight line as a course.

In this study, we extend our previous study. First, we verify that the same control program can trace an arc course. This is because a real course can be expressed in combination of straight and arc courses. Therefore, verifying the tracer for arc courses should be important to show applicability of model checking. Next, if the line tracer is not able to run along an arc course, we analyze turning angle using counter examples. Above-mentioned approaches will be useful to check performance properties in the design phase.

The roadmap of this paper is as follows. Sec. 2 outlines the foundations of our work and briefly describe our previous study. Sec. 3 show specification of a line tracer, and its implementation is described in Sec 4. Then, formal models used in verification are described in Sec. 5. Verification results are presented in Sec. 6, and Sec. 7 offers some discussion of these results. Finally, Sec. 8 provides a concluding summary and outline our future work.

2 PRELIMINARIES

In this section, we outline the background to our work and briefly show our previous study.

2.1 Model Checking

Model checking [4] is an automatic formal verification technique. Given a model that represents a system under consideration, and a logical formula that represents a property to be verified, model checking automatically determines whether or not the model satisfies a given property by exhaustively searching for the state space of the model. There are various kinds of model checking depending on expressive power of model and logical formula. In this study, we use timed automata to express models and Computation Tree Logic (CTL) for formulas. We use a model checker UPPAAL which takes timed automata as models and CTL formulas as property.

2.1.1 Timed Automata

A timed automaton is an extension of the conventional automaton with clock variables and constraints for expressing real-time dynamics. These are widely used in the modeling and analysis of real-time systems.

Definition 1 (constraints) We use the following constraints on clocks.

1. C represents a finite set of clocks.
2. Constraints $c(C)$ over clocks C are expressed as inequalities in the following BNF (Bacchus Naur Form).

$$E ::= x \sim a \mid x - y \sim b \mid E_1 \wedge E_2,$$

where $x, y \in C$, $\sim \in \{\leq, \geq, <, >, =\}$, and $a, b \in \mathbb{R}_{\geq 0}$, in which $\mathbb{R}_{\geq 0}$, is a set of all non-negative real numbers.

Time constraints are used to mark edges and nodes of the timed automata and for describing the guards and invariants.

Definition 2 (timed automaton) A timed automaton \mathcal{A} is a 6-tuple (A, L, l_0, C, T, I) , where

- A : a finite set of actions;
- L : a finite set of locations;
- $l_0 \in L$: an initial location;
- C : a finite set of clocks;
- $T \subseteq L \times c(C) \times A \times 2^C \times L$ is a set of transitions. The second and fourth items are called a guard and clock resets, respectively; and
- $I : L \rightarrow c(C)$ is a mapping from location to clock constraints, called a location invariant.

A transition $t = (l_1, g, a, r, l_2) \in T$ is denoted by $l_1 \xrightarrow{a, g, r} l_2$.

A map $v : C \rightarrow \mathbb{R}_{\geq 0}$, is called a clock assignment (or clock valuation). We define $(v + d)(x) = v(x) + d$ for $d \in \mathbb{R}_{\geq 0}$ and some $x \in C$.

For guards, resets and location invariants, we introduce some notation for clock valuations. For each guard $g \in c(C)$, a function $g(v)$ stands for the valuation of the guard expression g with the clock valuation v . For each reset r , where $r \in 2^C$, we introduce a function denoted by $r(v)$, and let $r(v) = v[x \mapsto 0], x \in r$. For each location invariant I , we shall introduce a function denoted by $I(l)(v)$, which stands for the valuation of the location invariant $I(l)$ of location l with the clock valuation v .

The dynamics of a timed automaton may be expressed via a set of states and their evaluations. Changes from one state to a new state may be as a result of either the firing of an action or an elapsed time.

Definition 3 (state of timed automaton) For a given timed automaton $\mathcal{A} = (A, L, l_0, C, T, I)$, let $S = L \times \mathbb{R}_{\geq 0}^C$ be the complete set of states of \mathcal{A} , where $\mathbb{R}_{\geq 0}^C$ is a complete set of clock evaluations on C .

The initial state of \mathcal{A} can be given as $(l_0, 0^C) \in S$. For a transition $l_1 \xrightarrow{a, g, r} l_2$, the following two transitions are semantically defined. The first one is called an action transition, while the latter one is called a delay transition.

$$\frac{l_1 \xrightarrow{a, g, r} l_2, g(v), I(l_2)(r(v))}{(l_1, v) \xrightarrow{a} (l_2, r(v))}, \quad \frac{\forall d' \leq d \ I(l_1)(v + d')}{(l_1, v) \xrightarrow{d} (l_1, v + d)}$$

The semantics of a timed automaton can be interpreted as a labeled transition system.

Definition 4 (semantics of a timed automaton) For a timed automaton $\mathcal{A} = (A, L, l_0, C, T, I)$, an infinite transition system is defined according to the semantics of \mathcal{A} , where the model begins with the initial state. By $\mathcal{T}(\mathcal{A}) = (S, s_0, \xrightarrow{\alpha})$, the semantic model of \mathcal{A} is denoted, where $\alpha \in A \cup \mathbb{R}_{\geq 0}$.

Definition 5 (run of a timed automaton) For a timed automaton \mathcal{A} , a run σ is finite or infinite sequence of transitions of $\mathcal{T}(\mathcal{A})$.

$$\sigma = (l_0, \nu_0) \xrightarrow{\alpha_1} (l_1, \nu_1) \xrightarrow{\alpha_2} (l_2, \nu_2) \xrightarrow{\alpha_3} \dots$$

2.1.2 Computation Tree Logic

In model checking, properties are written as logical formulas. Computation Tree Logic (CTL) [5] is a temporal logic suited to dealing with such formulas. Using CTL we are able to describe properties relating to behaviors of a program for a line tracer robot.

Let \mathcal{AP} be a set of atomic propositions. The syntax of CTL is defined as follows:

$$\begin{aligned} \varphi ::= & \perp \mid \top \mid p \mid \neg\varphi \mid \varphi \vee \varphi \mid \varphi \wedge \varphi \mid \varphi \rightarrow \varphi \\ & \mid \mathbf{AX}\varphi \mid \mathbf{EX}\varphi \mid \mathbf{A}\diamond\varphi \mid \mathbf{E}\diamond\varphi \mid \mathbf{A}\square\varphi \mid \mathbf{E}\square\varphi \\ & \mid \mathbf{A}[\varphi_1 \cup \varphi_2] \mid \mathbf{E}[\varphi_1 \cup \varphi_2], \end{aligned}$$

where p is an atomic proposition in \mathcal{AP} . The symbols \perp , \top , \neg , \vee , \wedge and \rightarrow have their usual meanings. The symbols \mathbf{X} (“next”), \diamond (“eventually”), \square (“globally”), and \cup (“until”) are temporal operators. The symbols \mathbf{A} (“always”) and \mathbf{E} (“exists”) are path quantifiers. Intuitively, temporal operators represent statements of a path, and path quantifiers represent statements on one or more paths which are branching forwards from a state. In a CTL formula, temporal operators are preceded by a path quantifier. Due to space limitation, we omit semantics. Please refer to Emerson [5] for details of the semantics of CTL.

For example, a safety property that “variable x is less than 10 for all paths” is written as a CTL formula $\mathbf{A}\square(x < 10)$.

2.1.3 UPPAAL

UPPAAL [2], is a popular model checker for extended timed automata. It supports model checking for both conventional and timed automata. UPPAAL allows verification of expressions described in an extended version of CTL. Note that, a property to be verified is called a query in the field of verification of timed automaton. Given a model and a query, UPPAAL checks whether or not the model satisfies the query. If the query does not hold, UPPAAL returns a counter example. A counter example is a run of the model, and presents sequence of locations that query does not hold. In addition,

UPPAAL supports local and global integers and primitive operations on integers, such as addition, subtract and multiplication with constants. Such expressions are also allowed on the guards of transitions. System models can be created from multiple timed automata which are synchronized via a CCS (Common-Channel Signaling)-like synchronization mechanisms. An important point is that, with the exception of clocks, the extended timed automaton used in UPPAAL cannot deal with real valued variables. We, therefore, have to round real values to integer values when we model the target systems.

2.2 Results from a Previous Study

In this subsection, we briefly mention about our previous study [6], [7]. The question at the core of our research is formal verification of embedded systems as hybrid systems. For that goal, our first step is verifying time-related properties of a real embedded application using UPPAAL. We set our application to a line tracing robot constructed by LEGO Mindstorms [8] with a control program written in Java base language LeJOS [9].

We presented verification of safety properties of the program for line tracing robot in terms of design verification. In the verification, we constructed two models expressed by timed automata, one for the control program and one for the motion control depending to the course. To construct these models, it is required that behaviors have to be modeled in discrete steps except for time clock. Sampling and quantization techniques are applied for the purpose. We showed experimental results of verification and presented model checking has power to check behaviors. We considered time-delay in the verification. However, the study was in preliminary stage because we set some limitations, such as no disturbance and handling only straight lines. Even limitations were set, we think that our previous study showed applicability of model checking for verifying real embedded systems.

3 SPECIFICATION

The whole system of line tracer consists of two parts; courses and a line tracer. We describe specifications in this section.

3.1 Course

For a line tracer robot, a *course* is a black line painted on white ground. Assume that course width w is constant. In general, a real course can be expressed in combination of straights and arcs. We verified that our control program can trace a straight course in our previous study. Therefore, we consider arc courses in this study. An arc course is expressed by radius r and central angle α .

3.2 Line Tracer

A *line tracer* is a vehicle which traces a course. In this study, we fix a line tracer that consists of a body, two motors, and two color sensors. Fig. 1 illustrates the relationships between constants and state variables.

A color sensor can discriminate colors. In this study, we assume that read value of the color sensor is two-valued, black

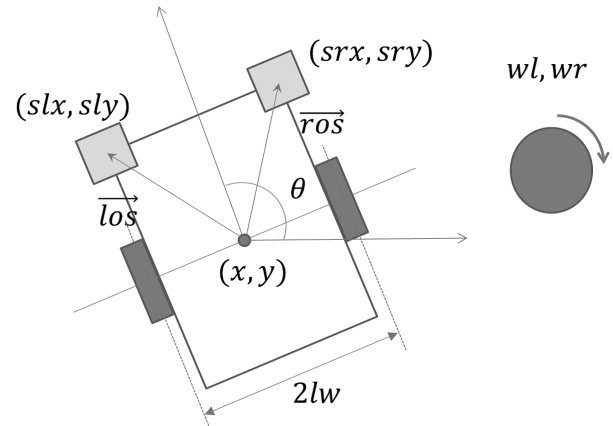


Figure 1: Constants and State Variables

Table 1: Logic for Color Sensors

		RightSensor	
		black	white
LeftSensor	black	go straight	turn left
	white	turn right	go straight

and white, by setting threshold. Then the line tracer reads colors of the course using two color sensors, and determines its motion by changing left or right wheel speeds. Table 1 shows the controller logic associated with read values of two color sensors. If, for example, the left sensor and the right sensor sense white and black respectively, then the line tracer will “turn right”. This is done by setting left wheel speed to high speed HS and right wheel speed to low speed LS . Note that there are delays in sensors and actuators, for example sleeping time before next sense-act loop and motor reaction.

4 IMPLEMENTATION

LEGO Mindstorms NXT [8] is a kit for assembling robots and machines with various actuators and sensors. The default programming language for LEGO NXT is Mindstorms, but there are other languages such as NXC (Not eXactly C) [10] and LeJOS [9] which supply various classes for NXT sensors and actuators. We use LeJOS for making the control program of the tracer. This is because Mindstorms is GUI base language and does not suit for modeling. Instead, LeJOS is Java based language and is easier to construct models from a program.

Fig. 2 shows our implemented controller program written in LeJOS. In this research, we use the same program used in our previous study mentioned in Sec. 2.2. Then, we try to verify that the program can trace arc courses.

5 MODEL

The line tracer system described in Sec. 3 is converted into two models; Controller model and Motion model. We introduce these models in this section,

```

import lejos.nxt.*;
public class Controller {
    public static void main(String[] args)
        throws Exception {
        int rid,lid;
        final int HS = 420, LS = 120, BLACK = 7,
        MS = 360, HSEC = 500;
        Color colorR ,colorL;
        ColorSensor sensorR =
            new ColorSensor(SensorPort.S3);
        // 1(S3):right
        ColorSensor sensorL =
            new ColorSensor(SensorPort.S4);
        // 2(S4):left
        Motor motor = new Motor();
        motor.B.setSpeed(MS);
        motor.C.setSpeed(MS);
        Thread.sleep(HSEC);
        // wait for devices to be stable
        motor.B.forward();
        motor.C.forward();
        while(true) {
            rid = sensorR.getColorID();
            lid = sensorL.getColorID();
            if (rid == BLACK)
                motor.B.setSpeed(LS);
            else
                motor.B.setSpeed(HS);
            if (lid == BLACK)
                motor.C.setSpeed(LS);
            else
                motor.C.setSpeed(HS);
            if (Button.readButtons()
                == Button.ENTER.getId())
                break;
        }
    }
}

```

Figure 2: Controller Program in LeJOS

Both Controller model and Motion model are expressed in timed automata. However, most of the state variables used in a line tracer have real values, and UPPAAL can only handle integer variables except for clock. Therefore, it is required to approximate state variables for discrete values to construct models in timed automata. We presented approximation of the state variables by applying sampling and quantization techniques in our previous study. In this study, we use the same models. Please refer papers [6], [7] for detail information of discretization techniques. Note that, we have modeled in the relative scale in this study. Therefore, units are not specified.

5.1 Controller Model

Controller model is a timed automaton which represents controller program for the line tracer. Fig. 3 shows Controller model which corresponds to the implemen-

Table 2: State Variables of a Line Tracer

variable	description
x :	x-coordinate of the center
y :	y-coordinate of the center
θ :	direction
$lsensor$:	sensed value of the left sensor
$rsensor$:	sensed value of the right sensor
s_{lx} :	x-coordinate of the left sensor
s_{ly} :	y-coordinate of the left sensor
s_{rx} :	x-coordinate of the right sensor
s_{ry} :	y-coordinate of the right sensor
wl :	revolution speed of the left wheel
wr :	revolution speed of the right wheel

tation in Sec. 4. Table 2 summarizes principle state variables in Controller model.

As described in Sec. 3, Controller decides motor speeds according to the four possible combinations of read values of the two color sensors. From the initial location represented as double circle, there are four transitions. Each of the transition corresponds to a pair of real value of sensors.

5.2 Motion Model

Motion model is a timed automaton which represents motions of the line tracer's coordinates of the gravity center and read values of color sensors. The line tracer keep on moving while the control program does not work because of delay or sleep time. Therefore, coordinates of center should be updated as independent of the Controller model to express behavior of the tracer. Fig. 4 shows the timed automaton which updates states variables at regular, discrete time intervals. The automaton of Motion model periodically calls functions `updateX`, `updateY`, `updateTheta`, `updateLSensor`, and `updateRSensor` which update state variables $x, y, \theta, lsensor$ and $rsensor$, respectively.

Read values of sensors depend on the coordinates of gravity center, the angle of the line tracer, and the course. Gravity center and angle are expressed by integer variables in this model. It is also required discretely handling of the course. There will be two methods for handling. First one is quantization, mapping the continuous course to discrete values. Second one is equation representation, the course is expressed in a formula. In this study, we adopt the second method. Let (s_x, s_y) be the coordinates of left or right sensor. Then, the read value of the sensor is decided to be black if (s_x, s_y) satisfies the following formula. Otherwise, the read value is decided to be white.

$$\left(r - \frac{w}{2}\right)^2 \leq s_x^2 + s_y^2 \wedge s_x^2 + s_y^2 \leq \left(r + \frac{w}{2}\right)^2$$

where r is radius of a circle course, and w is line width.

6 EXPERIMENTAL RESULTS

We obtain models representing the line trace system described in Sec. 5. In this section, we verify correctness of

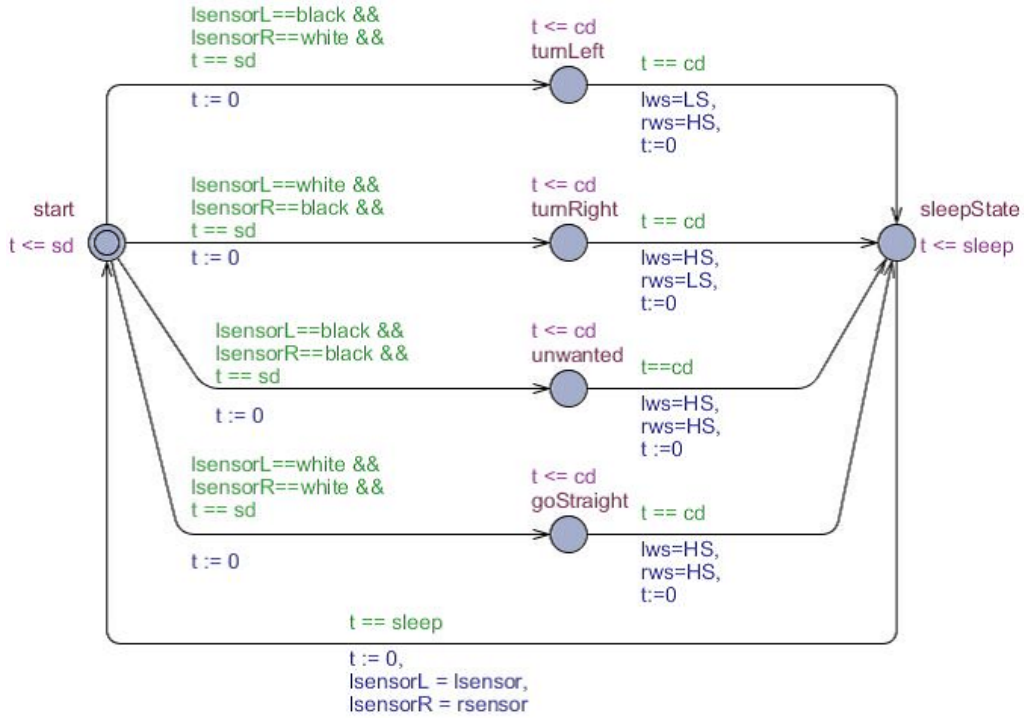


Figure 3: Controller Model

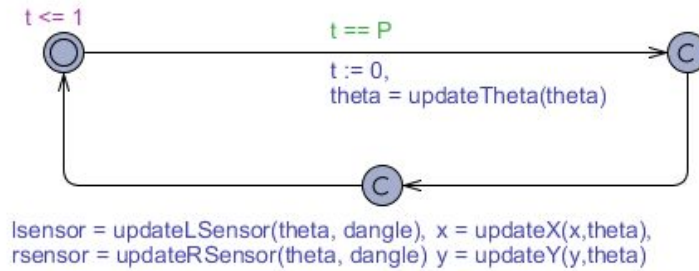


Figure 4: Motion Model

the control program. Verifications were performed using UP-PAAL 4.0.13 running on Windows 7 (64 bit), Intel Core i5-2400, 3.10GHz, with 8GB memory.

6.1 Verification of Specification

A line tracer is expected to trace a course. First, we verify whether or not the controller program satisfies the property. Therefore, what we need to verify is, i) the tracer runs along the course within a certain range, and ii) the tracer keeps on taking its route, i.e., does not get stuck.

To verify these requirements, we need to fix some initial values. Let initial values be as follows.

- coordinates of gravity center $(x, y) = (0, 0)$
- direction of the tracer $\theta = 90^\circ$
- width of the course $w = 100$

We then check the correctness of the line tracer by verifying the following queries.

1. $A \Box (firstquadrant \rightarrow inrange)$
where *firstquadrant* is $x \geq 0 \wedge y \geq 0$, *inrange* is $(r - \frac{w}{2} - ds)^2 \leq x^2 + y^2 \wedge x^2 + y^2 \leq (r + \frac{w}{2} + ds)^2$, and *ds* is a distance between center and a sensor, i.e., $ds = |\vec{r\oslash s}| = |\vec{l\oslash s}|$.
2. $E \Diamond (x < 0 \wedge y > 0)$

Query 1 represents that the gravity center of the tracer is always located within a certain range, $w/2 + ds$, from the line in the first quadrant. Note that we consider the gravity center (x, y) in this query, therefore *ds* is added to the allowable distance from the course. Here, target domain is limited to the first quadrant, because if the whole area is set to be a target, state explosion problem occurs. In addition, even if the target area is restricted, query 1 can not be verified because of the

state explosion problem.

To solve these problems, we slightly modified `Motion` model. We added a new location named `STOP` to `Motion` model. If the gravity center goes outside the first quadrant, then transit to the location `STOP`. This modification works on verification of query 1. Instead, we also have to modify query 1 considering the new location `STOP`. New query 1' is as follows.

$$A \square (\text{firstquadrant} \rightarrow \text{inrange} \vee \text{M.STOP})$$

where `M` is the variable name for `Motion` model in UPPAAL and `M.STOP` represents the location `STOP` in `Motion` model.

It is easily understand that the verification result for query 1' depends on the radius of the arc course. We verified query 1' by changing radius r . As a result, query 1' holds if $r \geq 277$ and does not hold if $r \leq 266$.

Query 2 is reachability checking that the line tracer eventually reaches to the second quadrant. This query is necessary to check behavior of the tracer, because query 1 only describe the distance from the course and does not describe movement. It makes no sense to check query 2 if query 1' does not hold. According to the above-mentioned results for query 1', we verified query 2 for $r \geq 277$. Then, we obtain verification results that query 2 holds for $r \geq 277$.

Ideally, conjunction of the two queries should be verified at once. Unfortunately, UPPAAL does not allow nesting of path quantifiers in a formula. Therefore we verified the queries one by one. However, when we consider both two queries together, it is possible to judge whether or not the tracer satisfies the specification. Note that, we verified dependency of radius by hand, but it is possible to be automated by generating UPPAAL model.

6.2 Analysis of Turning Angle

It is easily understand that verification results of query 1' depend on wheel speeds of the tracer. For example, if the tracer moves slowly, it will be able to keep on tracing longer. However, verification results of query 1' and query 2 do not describe distance from the initial position.

We calculate turning angle of the tracer by analyzing counter examples of query 1' for various wheel speeds. For that purpose, high wheel speed HS and low wheel speed LS are changed into $HS' = C_{ms}HS$ and $LS' = C_{ms}LS$ where C_{ms} is a coefficient. Then, we verify query 1' for some C_{ms} . When the query does not hold, we obtain a counter example which consists of a sequence of locations in evidence. UPPAAL has a function to generate the shortest trace as a counter example. By analyzing the counter example, it is possible to calculate the coordinate where the tracer turns off from the course. As an example, let radius r be fixed to 250. This is because that we know the tracer is not able to keeps on track in the first quadrant from the verification results in Sec. 6.1. Then, we think about intersection of the course and orbit of the tracer. Let the intersection be P , coordinates of before turning off be Q , and coordinates of after turning off be Q' . Then, P is an intersection of circle $x^2 + y^2 = (r \pm (w/2 + ds))^2$ and a line passing through Q and Q' .

Table 3: Speed Dependency and Turning Angle ($r = 250$)

C_{ms}	Q	Q'	P	α (deg)
1/2	—	—	—	$90 <$
2/3	(214, 428)	(213, 435)	(213.7, 429.8)	63.6
1	(262, 394)	(262, 405)	(262.0, 402.2)	56.9
2	(310, 360)	(314, 384)	(310.9, 365.7)	68.5
3	(310, 336)	(213, 396)	(310.9, 365.7)	68.5
4	(306, 306)	(213, 384)	(310.9, 365.7)	68.5

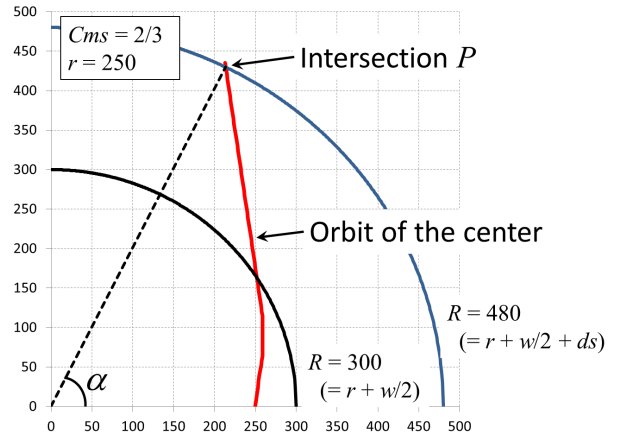


Figure 5: Orbit of the Tracer, Intersection, and Turning Angle

Table 3 shows C_{ms} , Q , Q' , P and α , where α (deg) is angle between x -axis and line passing through the origin and point P , obtained from the shortest counter examples. Note that there are no results for $C_{ms} = 1/2$ in Table 3, because query 1' holds. It is not surprisingly that verification results depend on wheel speeds. Query 1' holds for $C_{ms} = 1/2$ should be reasonable because this setting means slower move that arrows the tracer keeping on track. Fig. 5 shows a result of the orbit of the tracer obtained from the counter example, intersection P , and turning angle α for $C_{ms} = 2/3$ and $r = 250$. From the results except for $C_{ms} = 1/2$, central angle α is roughly constant. This result can be interpreted that angle α is the minimum turning radius for $r = 250$. This results seems natural, however, it indicates that model checking can be applied to analyze properties relating to turning angles.

7 DISCUSSION

In this section, we discuss our experiments and future work.

7.1 Discussion on the Experiments

We briefly return to our basic focus on our research question. We are motivated to know applicability of formal verification to real embedded systems, especially control continuous systems. Continuous systems are essentially hybrid systems, but we set our first target to verifying time-related properties. We also set another research question that we want to know applicability of verification techniques from the view point of design verification.

In this study, we divided the circle course into an arc course, the first quadrant, because of the state explosion problem. Here, we consider possibility of verification for tracing the entire route of the circle. To tackle this problem, straight-forward modeling seems unpromising according to the verification results in Sec. 6. To reduce the size of state space, one possibility is applying abstraction techniques such as data mapping and predicate abstraction. Another possibility will be combination of theorem proving and model checking.

Experimental results combined with our previous study, behavior of a line tracer is verified based on specification and a control program. We think our verification results indicate usefulness of model checking. However, there are still problems remained to verify real embedded system. One problem is scalability. Through our studies, parameters used in verification are not the same as those used in implementation and differ from LEGO Mindstorms kit in size. However, we believe that our parameter settings are acceptable to show applicability of model checking. The reason why we adjust parameters is the state explosion problem. If we set parameters as the same as real used values, the size of state space becomes too large, and model checker cannot respond in a reasonable time or it exhausts its available memory. This problem is widely known in the field of model checking.

Another problem is that we are not yet consider effects of errors and distributions. When we think of real embedded system, behaviors of the systems are disordered by disturbances or errors. It is natural that disturbances and error probabilistically occur. However, timed automaton is not suited for probabilistic event. Here, we give a little more thought to the tracer constructed by LEGO Mindstorms as a real embedded system. It is reported in [11] that motor speed of LEGO Mindstorms kit is approximately proportional to the parameter, but has error. Through this study, we have tried to handle errors associated with wheel speeds. We assumed that wheel speed includes a certain amount of error. If such error exists, errors are cumulated and make an impact on the position of the tracer. We confirmed that such errors affect to the result of verification. Unfortunately, we have not yet obtained systematic results.

7.2 Related Work

In this section, we briefly describe related work on formal verifications associated with control engineering.

One of similar researches is verification of real-time control program using UPPAAL [12]. In this paper, the authors constructed a brick sorter system using LEGO RCX and wrote control programs in Not Quite C (NQC). The paper presents verification of safety and liveness properties by automatic translation from the control program into UPPAAL models. Through the research, abstraction and reduction techniques are applied to construct discrete models from continuous systems. This approach is similar to ours, however, the brick sorter system is essentially a discrete system even though it contains time dependencies.

As with many control systems, a line tracer can be considered as a hybrid system by describing their movements using differential equations and their control programs in discrete

time. It is generally accepted that real embedded systems are too big to fully verify. Therefore, it is usual to focus on important behaviors. As an example of hybrid approaches, paper [13] described the verification of the behaviors of a line tracer by constructing a model using hybrid I/O automata and correctness proofs. In that paper, the authors presented verification of safety property, that is, a line tracer should move along a straight line and never run off. However, the authors noted that some time details, such as time delay between two motors, were not considered

In verification of robotics, a survey of model checking of the control system of robotics systems is reported [14]. In this survey, the authors summarize various techniques for verification and show verification of a robot control system. Safety and liveness properties are verified, but these properties were not related to continuous dynamics. Even though the survey does not cover the handling of continuous dynamics, it is a good resource. As a similar area, the verification of a real vehicle is presented [15]. Even though our aim is the verification of continuous systems, our approach in reflects those above, *i.e.*, conversion to timed automata using quantization and sampling.

8 CONCLUSION

In our previous study, we have verified that a line tracer runs along a straight line. In this research, we used the same control program for the tracer and showed the same models can keep track on arc courses. These are verified using UPPAAL with timed automata and logical formulas. We also presented that if the tracer cannot run from the first quadrant to the second quadrant, it is possible to calculate turning angle by analyzing counter examples.

We hope to extend this study to the analysis of more real embedded systems including disturbances and errors. To that purpose, expressive power of timed automaton is not sufficient as described in Sec. 7. We plan to express models in probabilistic timed automata (PTAs). We also intend to use the latest version of probabilistic model checker PRISM [16] which supports PTAs.

Another direction of future work includes a PID controller (proportional-integral-derivative controller), which is a widely used feedback control system. We used simple specification to control the line tracer, but PID control is widely used in control systems and control engineering. When PID control is applied to a line tracer, it enables smooth motion. However, PID control is essentially hybrid system, which continuous and discrete dynamics are mixed with time progression. Several approaches have been proposed to handle hybrid systems. One of these approach is hybrid automata [17] which is a formal model for describing discrete-continuous systems.

ACKNOWLEDGMENTS

Part of this research is supported by the Telecommunications Advancement Foundation (TAF).

REFERENCES

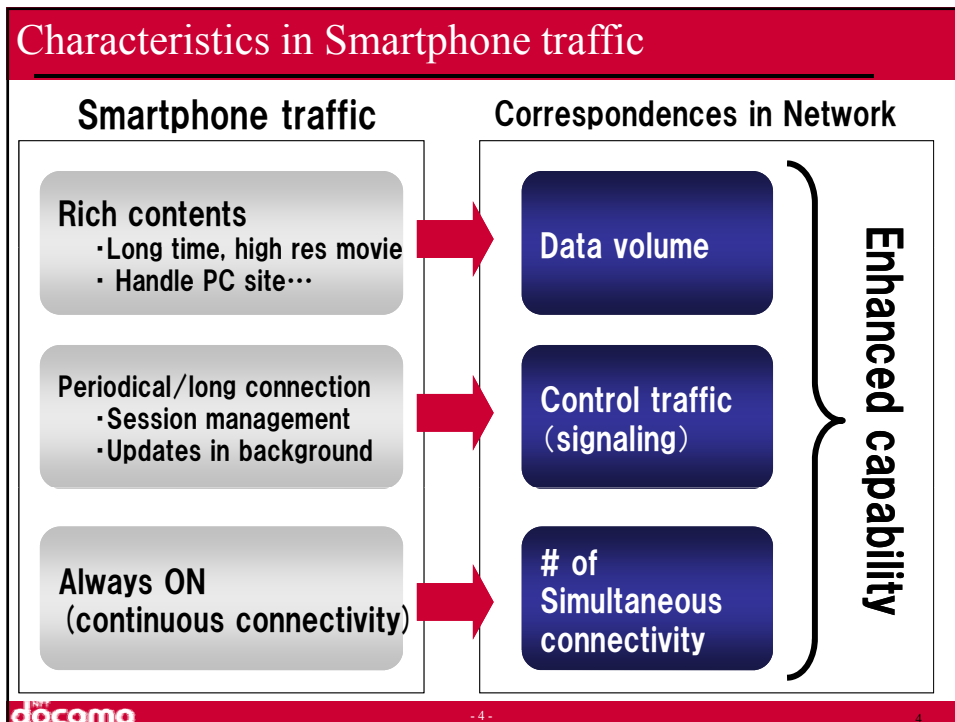
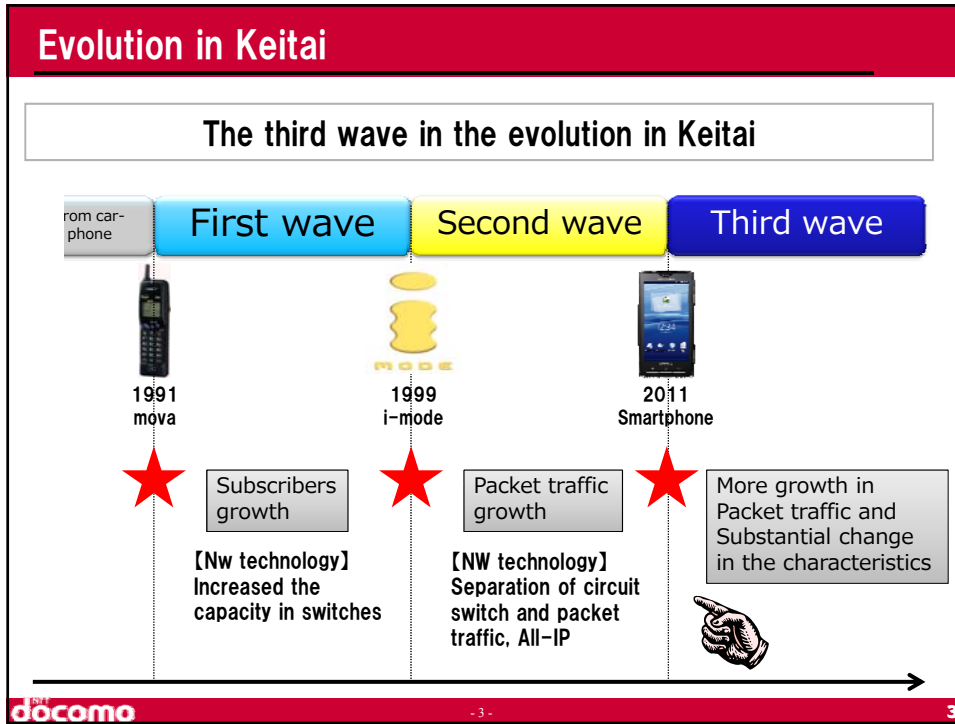
- [1] Rajeev Alur and David L. Dill. A theory of timed automata. *Theoretical Computer Science*, 126(2):183–235, 1994.
- [2] Johan Bengtsson and Wang Yi. Timed automata: Semantics, algorithms and tools. In Jorg Desel, Wolfgang Reisig, and Grzegorz Rozenberg, editors, *Lectures on Concurrency and Petri Nets, Advances in Petri Nets (4th ACPN'03)*, volume 3098 of *Lecture Notes in Computer Science (LNCS)*, pages 87–124. Springer-Verlag (New York), Eichstatt, Germany, September 2003, selected revised paper 2004.
- [3] A. Schild and J. Lunze. Control design by means of embedded maps. In J. Lunze and F. Lamnabhi-Lagarrigue, editors, *Handbook of Hybrid Systems Control*, chapter 6.5, pages 231–247. Cambridge University Press, 2009.
- [4] Edmund M. Clarke, Orna Grumberg, and Doron Peled. *Model Checking*. MIT Press, 1999. To appear.
- [5] E. Allen Emerson. Temporal and modal logic. In *Handbook of Theoretical Computer Science*, volume B, chapter 16, pages 995–1072. Elsevier, 1990.
- [6] Kozo Okano, Toshifusa Sekizawa, Hiroaki Shimba, Hideki Kawai, Kentaro Hanada, Yukihiro Sasaki, and Shinji Kusumoto. Verification of safety property of line tracer program using timed automaton model. In *International Workshop on Informatics (IWIN2012)*, pages 136–142, 2012. Chamonix Mont-Blanc, France.
- [7] Kozo Okano, Toshifusa Sekizawa, Hiroaki Shimba, Hideki Kawai, Kentaro Hanada, Yukihiro Sasaki, and Shinji Kusumoto. Verification of safety properties of a program for line tracing robot using a timed automaton model. *Special Issue of the International Journal of Informatics Society (IJIS)*. (to appear).
- [8] LEGO Mindstorms NXT official website. <http://www.legoeducation.jp/mindstorms/>.
- [9] LeJOS Java for LEGO Mindstorms. <http://lejos.sourceforge.net>.
- [10] NXC Tutorial. <http://bricxcc.sourceforge.net/nbc/nxcdoc/NXCtutorial.pdf>.
- [11] Kazuki Yamabe. Measurement of performance characteristic of LEGO NXT using LeJOS, 2011. (undergraduate thesis in Osaka Gakuin University, written in Japanese).
- [12] Torsten K. Iversen, Kare J. Kristoffersen, Kim G. Larsen, Morten Laursen, Rune G. Madsen, Steffen K. Mortensen, Paul Pettersson, and Chris B. Thomasen. Model-checking real-time control programs - verifying lego mindstorms systems using uppaal. In *In Proc. of 12th Euromicro Conference on Real-Time Systems*, pages 147–155. IEEE Computer Society Press.
- [13] A. Fehnker, F. W. Vaandrager, and M. Zhang. Modeling and verifying a Lego car using hybrid I/O automata. In *Models, Algebras, and Logic of Engineering Software*, volume 191 of *NATO ASI Series III*, pages 385–402. IOS Press, 2003.
- [14] Natasha Sharygina, James Browne, Fei Xie, and Vladimir Levin. Lessons learned from model checking a nasa robot controller. In *Formal Methods in Systems Design Journal*, 2004.
- [15] Martin Proetzsch, Karsten Berns, T. Schuele, and K. Schneider. Formal verification of safety behaviours of the outdoor robot raven. In *Fourth International Conference on Informatics in Control, Automation and Robotics (ICINCO)*. INSTICC Press, 2007.
- [16] M. Kwiatkowska, G. Norman, and D. Parker. PRISM 4.0: Verification of probabilistic real-time systems. In G. Gopalakrishnan and S. Qadeer, editors, *Proc. 23rd International Conference on Computer Aided Verification (CAV'11)*, volume 6806 of *LNCS*, pages 585–591. Springer, 2011.
- [17] Thomas A. Henzinger. The theory of hybrid automata. pages 278–292. IEEE Computer Society Press, 1996.

Keynote Speech :
Mr. Hiroshi Inamura
(Research Laboratories,
NTT DOCOMO, Inc.)

Research Activities in DOCOMO
~ Toward Smart Life Partner ~
Hiroshi Inamura, Research Labs.



Use of data for insight of NOW



Our research on power efficiency...

Topic: Battery longevity in smart phone

- Issue: apps runs Tasks in background while screen turned off, that causes extra energy consumption

	Dissatisfaction in smart phone	Ratio (%)
1	Short battery life	68.5
2	Usability in touch panel	33.3
3	No water proof	25.1
4	Expensive and bad usability	23.4

2011 Aug. Macromill Survey

While screen is off

☐ Feature phone

☐ Smartphone

➔ Looking into the data to see the behavior of the back ground tasks in Android

- 5 -

Unexpected Traffic on Screen Turned On

- Wakeup clock on Android

IP communication starts on the Alarm Clock Ring

Apps updates info with server on Screen ON
Making connection at 8:00

Screen OFF

➔

Turned On
In morning

Alarm Screen

➔

User operation

Home Screen

- 6 -

Global Synchronization

- If many terminal start sending packets at the same time,
 - **Global Synchronization : Traffic Concentration**
 - ✓ It may overload routers and switches in network

Many terminals
Sending packets exactly the same
time

Traffic
concentration

Packet transmission

Time

docomo - 7 -

Android™ Appli Development Guideline ~For Efficient Communication Control~

- Describes appli implementation methods to achieve balanced improvement of application responsiveness and battery consumption
- Released on October 18, 2012
- The Guideline explains how to implement applications to improve application responsiveness and battery consumption while keeping their balance and provides background knowledge on mobile networks and specific improvement methods.

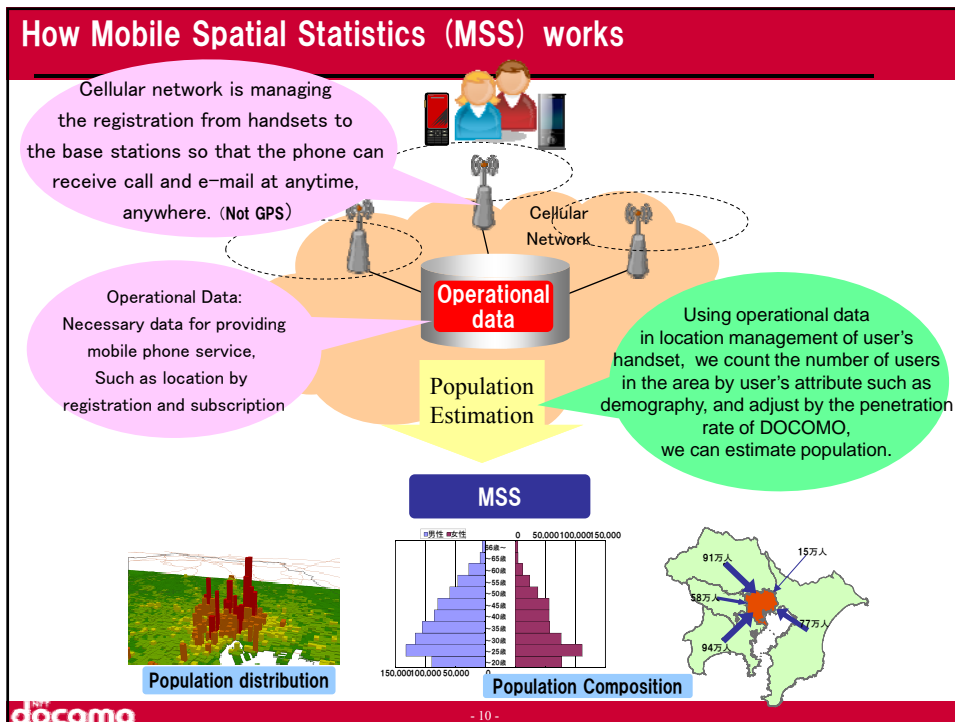
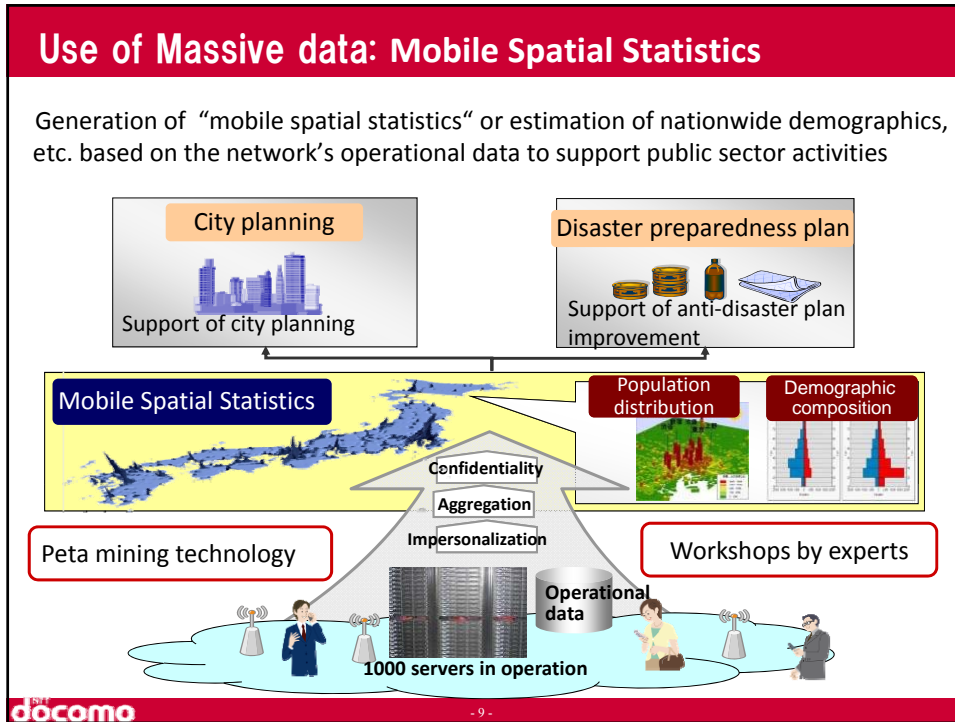
Traffic
concentration

目次	
1. 概要	1
1.1 概要	1
1.2 本書の目的	2
1.3 本書の構成	3
2. 本書	4
2.1 本書の目的	4
2.2 本書の構成	5
2.3 本書の更新履歴	6
2.4 本書の著作権	7
2.5 本書の免責事項	8
2.6 本書の印刷	9
2.7 本書の発行	10
2.8 本書の印刷	11
2.9 本書の発行	12
2.10 本書の印刷	13
2.11 本書の発行	14
2.12 本書の印刷	15
2.13 本書の発行	16
2.14 本書の印刷	17
2.15 本書の発行	18
2.16 本書の印刷	19
2.17 本書の発行	20
2.18 本書の印刷	21
2.19 本書の発行	22
2.20 本書の印刷	23
2.21 本書の発行	24
2.22 本書の印刷	25
2.23 本書の発行	26
2.24 本書の印刷	27
2.25 本書の発行	28
2.26 本書の印刷	29
2.27 本書の発行	30
2.28 本書の印刷	31
2.29 本書の発行	32
2.30 本書の印刷	33
2.31 本書の発行	34
2.32 本書の印刷	35
2.33 本書の発行	36
2.34 本書の印刷	37
2.35 本書の発行	38
2.36 本書の印刷	39
2.37 本書の発行	40
2.38 本書の印刷	41
2.39 本書の発行	42
2.40 本書の印刷	43
2.41 本書の発行	44
2.42 本書の印刷	45
2.43 本書の発行	46
2.44 本書の印刷	47
2.45 本書の発行	48
2.46 本書の印刷	49
2.47 本書の発行	50
2.48 本書の印刷	51
2.49 本書の発行	52
2.50 本書の印刷	53
2.51 本書の発行	54
2.52 本書の印刷	55
2.53 本書の発行	56
2.54 本書の印刷	57
2.55 本書の発行	58
2.56 本書の印刷	59
2.57 本書の発行	60
2.58 本書の印刷	61
2.59 本書の発行	62
2.60 本書の印刷	63
2.61 本書の発行	64
2.62 本書の印刷	65
2.63 本書の発行	66
2.64 本書の印刷	67
2.65 本書の発行	68
2.66 本書の印刷	69
2.67 本書の発行	70
2.68 本書の印刷	71
2.69 本書の発行	72
2.70 本書の印刷	73
2.71 本書の発行	74
2.72 本書の印刷	75
2.73 本書の発行	76
2.74 本書の印刷	77
2.75 本書の発行	78
2.76 本書の印刷	79
2.77 本書の発行	80
2.78 本書の印刷	81
2.79 本書の発行	82
2.80 本書の印刷	83
2.81 本書の発行	84
2.82 本書の印刷	85
2.83 本書の発行	86
2.84 本書の印刷	87
2.85 本書の発行	88
2.86 本書の印刷	89
2.87 本書の発行	90
2.88 本書の印刷	91
2.89 本書の発行	92
2.90 本書の印刷	93
2.91 本書の発行	94
2.92 本書の印刷	95
2.93 本書の発行	96
2.94 本書の印刷	97
2.95 本書の発行	98
2.96 本書の印刷	99
2.97 本書の発行	100
2.98 本書の印刷	101
2.99 本書の発行	102
2.100 本書の印刷	103

docomo

http://www.nttdocomo.co.jp/service/developer/smart_phone/technical_info/etc/

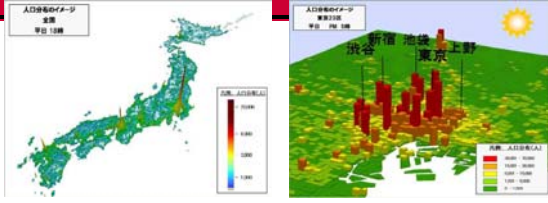
8



Example of Mobile Spatial Statistics

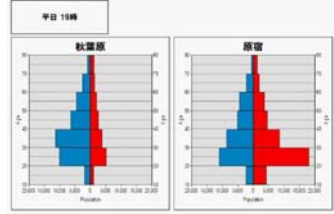
Population distribution

Population distribution in the Area
Distribution in Japan
Distribution in Tokyo metro
⇒ **Geographical Distribution**

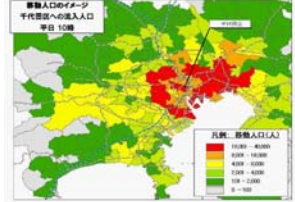


Population Composition

Sex/Age/Dwelling area-wise stats
Akihabara area / Harajuku area
⇒ **Population stats by attributes**



Non resident population in
Tokyo Chiyoda area
⇒ **Population in dwelling area**



docomo
- 11 -
1.1

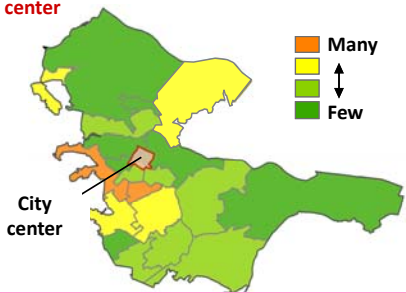
Mobile Spatial Statistics: application in public sector

Effectiveness of mobile spatial statistics in the public sector confirmed in joint research projects

City planning

(Research objective) To estimate the demand for public transportation to the city center

Number of visitors from each area to the city center

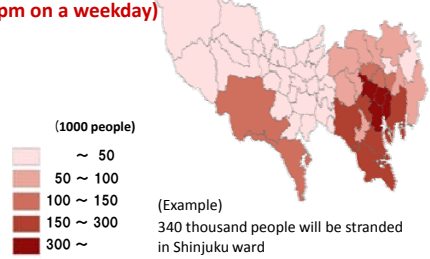


Useful as fundamental data for future discussions on public transportation

Disaster preparedness plan

(Research objective) To estimate the number of stranded people in Tokyo when hit by an epicentral earthquake

4.25 million people will be stranded in entire Tokyo are (worst case scenario: hit by major earthquake at 3.pm on a weekday)



Useful for discussions on measures to support stranded people

docomo
- 12 -

Vision and Research Activity For Near Future



- 13 -

Global changes and challenges
Accelerating globalization
Serious environmental issues
Quality of personal communication changing

HEART

Harmonize Social contribution beyond borders, across generations

Evolve Evolution of service and network

Advance Advance industries through convergence of services

Relate Creating joy through connections

Trust Support for safe, secure, and comfortable living

Mobile—part of the fabric of life

Challenging the Mobile Frontier
MAGIC

- M Mobile Multimedia
- A Anytime, Anywhere, Anyone
- G Global Mobility Support
- I Integrated Wireless Solution
- C Customized Personal Service

2013 NTT DOCOMO INC. All Rights Reserved

Shaping a Smart Life

Aim to bring smart life into reality by propelling the evolution of mobile services and new value creation through convergence of industries/services leveraging DOCOMO's clouds

DOCOMO's clouds

"Personal" cloud
Platform underpinning a wide range of services for consumers

"Business" cloud
Solutions platform for provision of new business styles

Network cloud
Platform that adds value through sophisticated information and communication processing performed on the network

- 15 -

Glass-type Wearable Device

- A futuristic mobile phone: a wearable terminal with all functions on its glass-type device
- Numerous sensors embedded in the device enable hands-free videophone, virtual office, vital monitoring, etc.
- A thin-client terminal composed of a human interface device (HID) and radio equipment alone; data analysis and accumulation are performed on the cloud.

Hands-free videophone

Virtual office

AR shopping

Vital monitoring

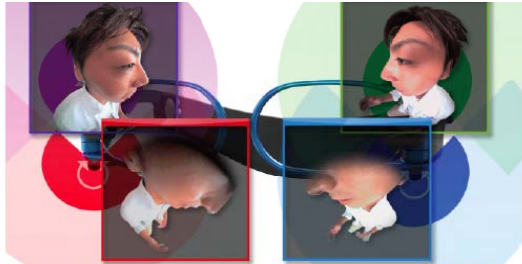
BioSensors
Biological Gas
Body Temperature
Blood Pressure
Pulse Wave

- 16 -
Copyright (c) 2013 NTT DOCOMO INC. All Rights Reserved.


Hands-Free Videophone

- Wearing these glasses, you can talk over the videophone while showing the other person your front face image. You do not have to keep holding the phone while talking.
- You can “take a photo of yourself” through multiple super-wide angle cameras set on the glasses frame.
- This device can eliminate some issues with conventional videophones such as “you need to keep holding the camera while talking “ and “the person shown on the screen always look at somewhere off-center.”

Images of the face and the body of the user wearing the glass-type terminal captured by cameras on the terminal.




Glass-type terminal Camera to shoot the background






Super-wide angle camera (to shoot the user wearing the glass-type terminal)

A user wearing the glass-type terminal is taking on the videophone.




Composite Image



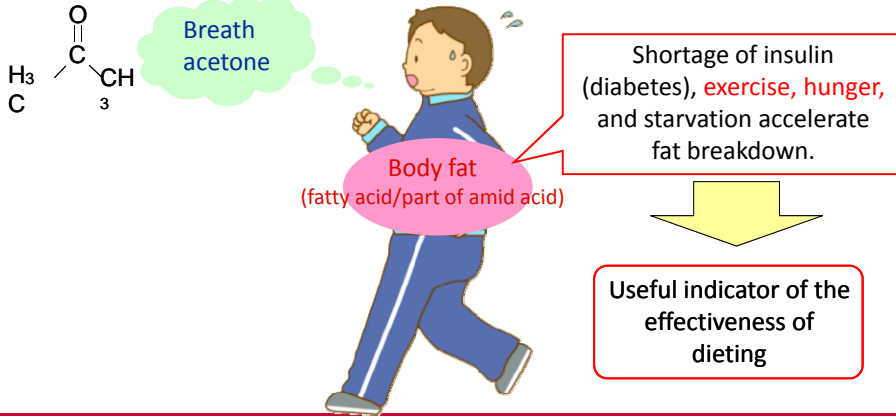

- 17 -




What is Breath Acetone?



- Acetone is generated in the blood as a metabolic product of fat breakdown.
- Acetone is released out of the body through the lungs as a component of breath gases.
- It is difficult to detect Acetone due to its low concentration (normal average concentration: 0.5ppm = 0.00005%)

$$\begin{array}{c}
 \text{O} \\
 \parallel \\
 \text{H}_3\text{C}-\text{C}-\text{CH}_3 \\
 | \\
 \text{C}
 \end{array}$$




- 18 -


Diet Support Using Breath Measurement docomo





はらぺこメータ



脂肪燃焼メータ

図3 呼吸アセトン計測結果のスマートフォン上での表示例

- Exhibited at CEATEC Japan 2011


docomo
- 19 -
Copyright (c) 2013 NTT DOCOMO INC. All Rights Reserved.

Breath Measurement Device docomo


Patented Technology

Calculation of breath acetone concentration through signal processing using two types of semiconductor-based bio-sensors with different gaseous sensitivity characteristics without gas separation

World's smallest gas analyzer (exact calculation using columns performing gas separation)

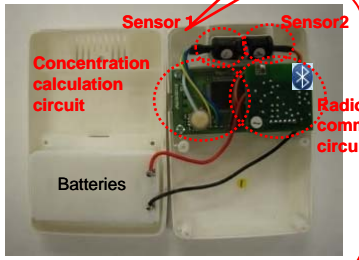


Conventional device



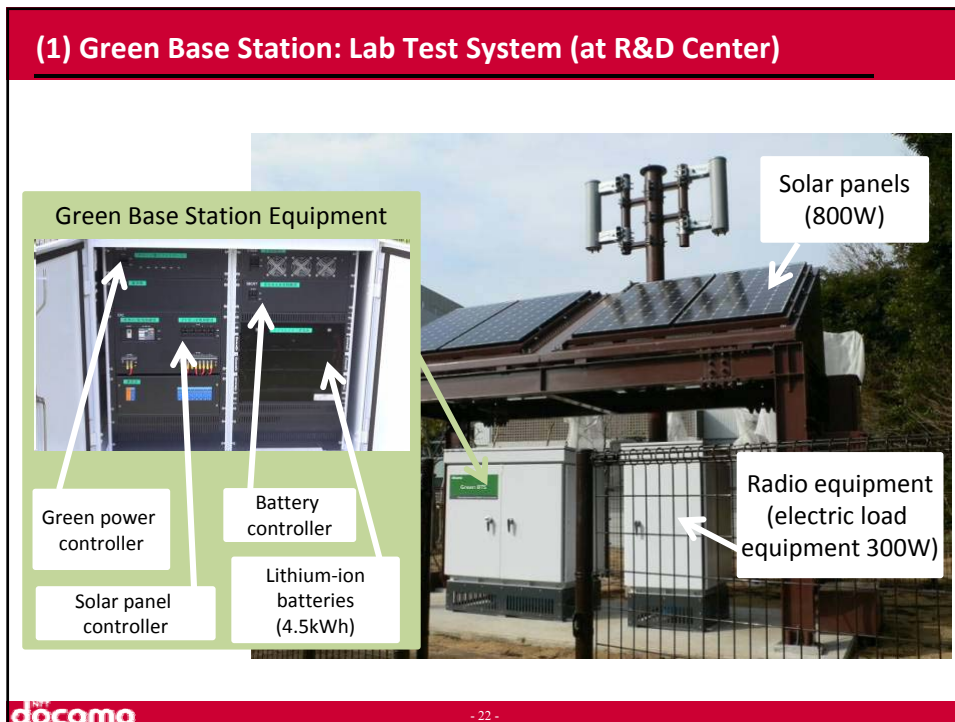
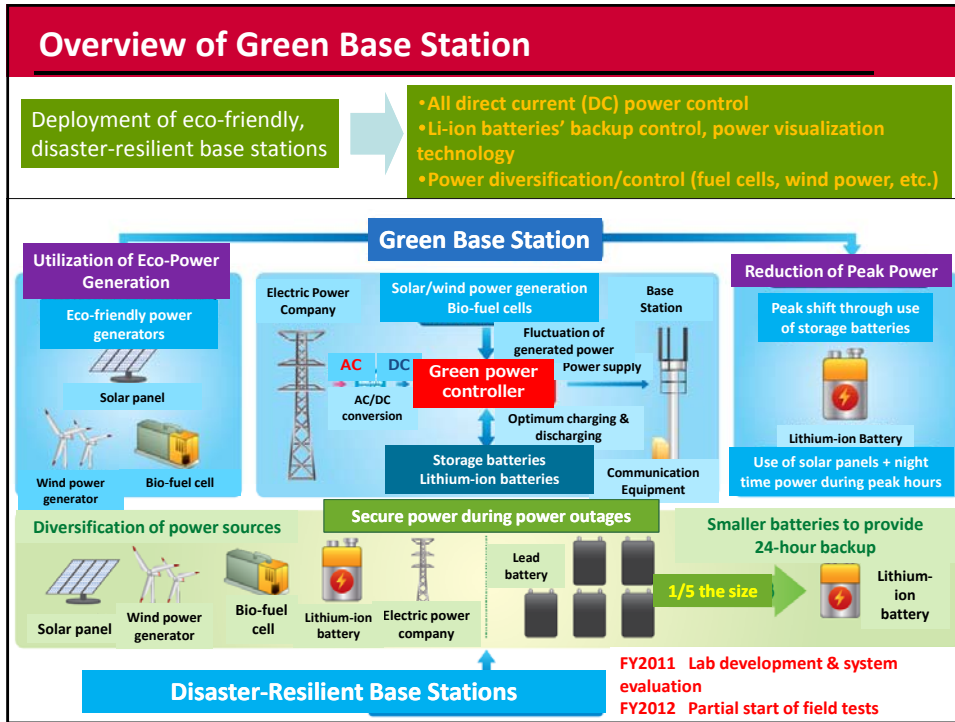
Our device

Weight reduced from 6kg to 125g. Its lightweight & compactness enables anytime anywhere measurement



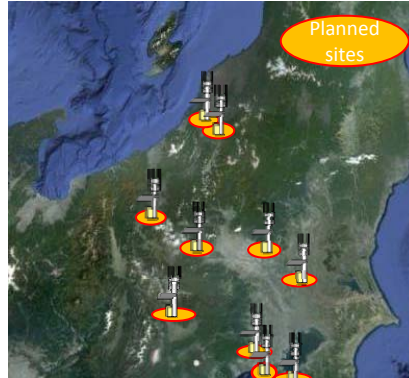
(Case size)
65 × 100 × 25mm, 125g (with batteries)
(Power consumption)
DC 3V, 100mA or less, approx. 10-hour operation time with 2 AA batteries

docomo
- 20 -
Copyright (c) 2013 NTT DOCOMO INC. All Rights Reserved.



(2) Field Test (started in March 2013)

Field test is underway in 3 separate regions: the Pacific side, the Sea of Japan side, and the inland (3 base stations installed, others under construction)



More utilization of solar energy and smart grid
Power reallocation between base stations
Deployment of a smart grid to Green Base Stations

Thank you

Session 5:

Network 2

(Chair: Yoshitaka Nakamura)

Improving of Terminal-Independent Handover Method with SIP Mobility

Yoshio Oda^{*}, Yoshitaka Nakamura^{**}, and Osamu Takahashi^{**}

^{*}Graduate school of Systems Information Science, Future University Hakodate, Japan

^{**}School of Systems Information Science, Future University Hakodate, Japan
{g2112011, y-nakamr, osamu}@fun.ac.jp

Abstract - With the development of mobile communication technology in public wireless LAN, handover technologies become more important. The purpose of this study is to improve the communication performance of the terminal-independent handover method, and to achieve handover between different types of network. We propose the method to implement the expanded SIP Mobility to wireless LAN routers, and the method to shorten the acquisition time of the IP address by dynamically changing the transmission timing of the router. And we evaluate these proposed methods by the simulation experiments.

Keywords: Handover, SIP, Wireless LAN.

1 INTRODUCTION

Recently, IPv4 addresses are in danger of depletion, and the use of IPv6 addresses is promoted. However, the incompatibility between IPv4 and IPv6 has prevented the wide use of IPv6 address. IPv6 shift test [1] is carried out on a global scale in 2011, and it is expected that use environment of the IPv6 is regulated well.

There are also significant developments in the use of public wireless LAN and mobile terminals such as note PCs and smart phones. Therefore, the opportunities for communicating terminal to change communicating wireless LAN router by its movement are expected to increase. However, in the current public wireless LAN environment when a communicating mobile terminal moves over multiple domain of different wireless LAN router, the IP address of the mobile terminal changes. When an IP address is changed, the TCP which is transport protocol cannot maintain the connection, and disconnection of communication is occurred. Therefore, the importance of handover techniques which keep communication when the mobile terminals move between wireless LAN routers has been increasing. There are some researches of handover techniques such as a PMIPv6 (Proxy Mobile IPv6) [2] and the method using a SIP (Session Initiation Protocol) Mobility method [3, 4].

The PMIPv6 is one of the terminal-independent handover methods. This technology is consisted of the LMA (Local Mobility Anchor) and the default routers. The LMA is the position management server of the mobile terminals. The PMIPv6 switches over the communication path between the LMA and the default routers during the handover of the mobile terminal. Therefore, even mobile terminals can use the PMIPv6 without the need for any special features. However, communication passes is through the LMA. As a

result, the LMA is predisposed toward bottleneck when the number of communicating mobile terminals on the increase.

SIP runs sessions establishing, changing, and cutting between terminals. The SIP Mobility is session changing. Terminals can have a communication path of choice after SIP session consisting. Therefore, communication is insulated from the influence of bottlenecks. SIP Mobility enables handovers when it is implemented in the mobile terminals. However, every mobile terminal cannot run a handover with this method because it presupposes implementation to the mobile terminal.

We propose a handover method that uses SIP Mobility expansion. This method involves introducing the SIP mobility expansion into the ingress router. The router can establish and change the SIP session. The communication does not pass through a specific server, and the mobile terminal can use this technology without the need to have any special feature. However, the number of mobile terminals becomes depleted by a handover because the mobile terminals that take this IP address are long. We focused on the RA (Router Advertisement) packet, which is sent by a wireless LAN router. The RA packet assigns IP addresses to the mobile terminal. We propose dynamically changing the sending timing of the RA packet and thus reducing the amount of time needed for the mobile terminal to receive IP addresses and thus improve communication and present the evaluation result of the simulation.

2 RELATED TECHNOLOGY

This chapter is an overview of the PMIPv6 and the SIP Mobility.

2.1 PMIPv6

The PMIPv6 is a terminal-independent handover method in which the mobile terminal runs a handover without the need for any special features and can hold the TCP connection. This technology consists of the LMA and the default router. It constructs a tunnel between the LMA and the default router, and the terminal's communication path is controlled. IPv6 consists of a Network Prefix and Link Local Address. The Network Prefix is a network identifier, and terminals are assigned this by the network. The Link Local Address is derived from the MAC address of terminals. The LMA keeps tabs on the Network Prefix and identifier of the mobile terminal, and the mobile terminal assigns the same Network Prefix to the LMA after a handover. Therefore, the mobile terminal can keep up the

same IP address even though it runs a handover. However, the mobile terminal cannot run a handover between MAGs that connect other LMA. Therefore, the mobile terminal can use the PMIPv6 only if the mobile terminal runs a handover between MAGs that have the same connection as the LMA. The processing of the PMIPv6 is shown in Figure 1.

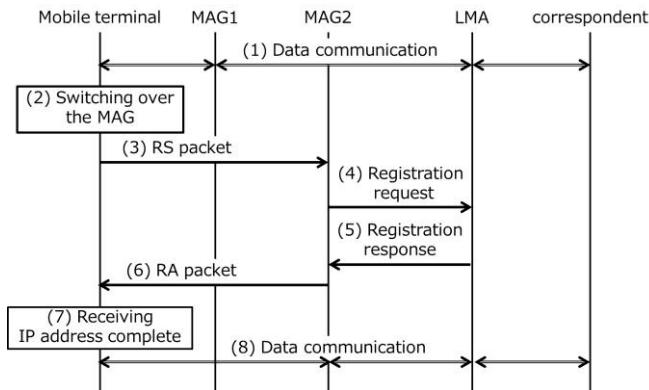


Figure 1: Processing of PMIPv6 handover

This process premises on connecting the mobile terminal to the MAG 1, and the mobile terminal communicates to a correspondent. A tunnel is constructed between the LMA and the MAG 1, and data communication is run by way of the LMA (Figure 1 (1)). The mobile terminal sends a RS (Router Solicitation) packet to the MAG 2 when it switches over to MAG (Figure 1 (2) - (3)). A RS packet is a required of a RA packet. The MAG 2 sends the information of the mobile terminal to the LMA when the MAG 2 receives the RS packet (Figure 1 (4)). The LMA updates the information of the mobile terminal, and it sends an acknowledgement to the MAG 2 (Figure 1 (5)). The MAG 2 constructs the LMA and sends the RA packet to the mobile terminal (Figure 1 (6) - (7)). The mobile terminal resumes communication and completes a handover (Figure 1(8)).

The mobile terminal communicates by way of the LMA with the PMIPv6 and uses the PMIPv6 without the need for any special features because a tunnel switches over between the LMA and MAGs.

2.2 SIP Mobility

The SIP is a session initiation protocol and runs a SIP session establishing, changing, and cutting between more than one terminal. The SIP Mobility is the SIP session changing. The SIP can establish a SIP session that only terminal handles and the SIP message is sent and received by way of the SIP server. After establishing the SIP session, mobile terminals can communicate without the SIP server. The mobile terminal describes the IP address that gives out the destination wireless LAN router to the SIP message, and re-establishes the SIP session with a correspondent, and is able to switch over to communication path. The processing of the SIP Mobility is shown in Figure 2.

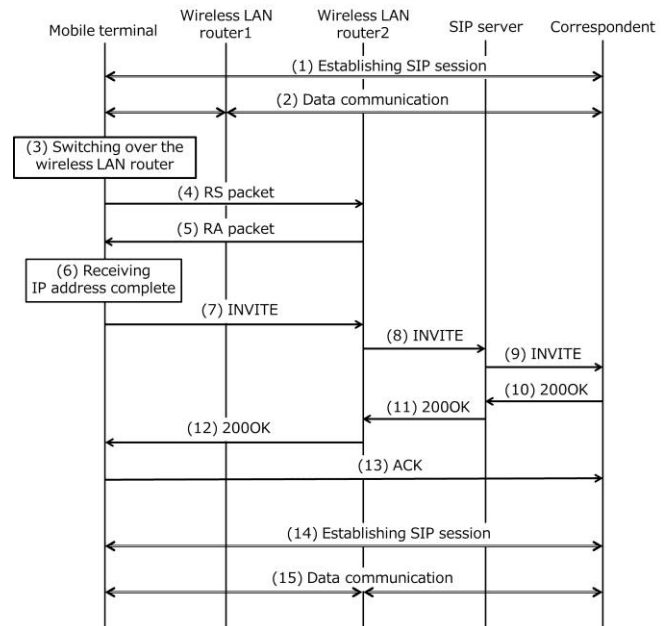


Figure 2: Processing of SIP Mobility handover

This process premises on connecting the mobile terminal to wireless LAN router 1, and the mobile terminal and correspondent handle the SIP (Figure 2 (1)). They establish the SIP session and communicate data (Figure 2 (2)). The mobile terminal sends a RS packet to wireless LAN router 2 when it switches over to wireless LAN routers, and wireless LAN router 2 sends a RA packet to the mobile terminal (Figure 2 (4) - (5)). The mobile terminal acquires the IP address after receiving the RA packet (Figure2 (6)). The mobile terminal sends this IP address to the correspondent and re-establishes the SIP session (Figure 2 (7) - (14)). Then, they restart data communication (Figure 2 (15)).

The mobile terminal can maintain data communication because it re-establishes the SIP session. A portion of the terminal converges into a load because data communication is done by way of SIP server.

2.3 Tasks

There are five problems with the existing method

- i. Bottleneck-prone
The PMIPv6 impedes data communication because this technology gets centered on data communication to the LMA, where bottlenecks tend to occur.
- ii. Technology not available in a different network
The mobile terminal cannot run a handover with the PMIPv6 in a different network because this technology can only be used between MAGs connecting the same LMA.
- iii. Unsustainable TCP connection
The mobile terminal cannot run a handover as the TCP connection is unsustainable with the SIP Mobility because the mobile terminal changes the IP address.

iv. Need for specific features for the mobile terminal

The SIP Mobility must be implemented in the mobile terminal.

v. Disconnection for an amount of time during handover

The PMIPv6 and the SIP Mobility take long to get the IP address, disconnect for an amount of time, and proceed with communication after a handover with the mobile terminal because they do not implement any specific features to the mobile terminal. As a result, we thought they impede communication performance.

3 PROPOSED METHOD

This chapter is an overview of the proposed method. Details on the system are given in our past research [5]. Therefore we expound shortening the time of taking IP address because we are not expound research past of ours.

3.1 Overview

In this study, we propose a terminal-independent handover method in which the mobile terminal uses a handover technology without the need for any special features and accepts a handover in a different network. The SIP Mobility no longer needs to be the communication path by way of SIP server. We focus on the SIP Mobility for problems i, ii, iii, and iv in Section 2.3. Routers are implemented in the SIP, which establishes and changes the SIP session between routers and supports the handover of a mobile terminal. Therefore, the mobile terminal uses our proposed method without the need for any special features. This method can run a handover in a different network by setting up a SIP server in internet. Only one SIP cannot maintain the IP address of the mobile terminal. Therefore, wireless LAN routers give out the same Network Prefix for all mobile terminals. The mobile terminal does not change the IP address even if it connects to every wireless LAN route, and it can maintain TCP connection. We call this method using expanded SIP Mobility method.

We explain problem v of Section 2.3. The most part of disconnection time during a handover is said to be the time spent receiving a RA packet to the mobile terminal after finishing off the L2HO (Layer 2 handover) and the time spent on DAD (Duplicate Address Detection) [6]. We measured the time between disconnection and reconnection to wireless LAN routers for the mobile terminal, and this time was determined to be about 3.8 seconds. The time spent for DAD was about 1 second. Therefore, the mobile terminal took over 2 seconds to receiving the RA packet. Therefore, we propose improving the communication performance by means of shortening the time spent on taking the IP address.

3.2 Prerequisite

The proposal deals with the possibility of the IPv6 network. An example of the network configuration is shown in Figure 3.

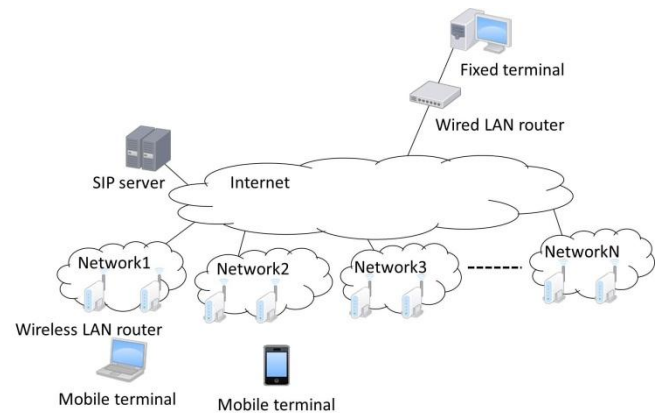


Figure 3: Network configuration example

The mobile terminal moves between network1 and network N, and this network is a different ISP.

3.1 Using expanded SIP Mobility method

A handover method that uses SIP Mobility expansion is a terminal-independent handover method. This method is a SIP to introduce and add function for a handover to the default routers of terminals. The SIP Mobility expansion has function of sending the identifier from the mobile terminal to the SIP server, sends passage of the mobile terminal to other routers, and establishes SIP session between routers. This method expands the SIP server to have the information that manages the mobile terminals. This information is connecting wireless LAN routers and correspondents in the mobile terminal, thus expanding the functions of routers to giving out the same Network Prefix for all mobile terminals, running encapsulation and decapsulation for packets and the detection of the passage of mobile terminals, and taking the information from the SIP server.

When the mobile terminal communicates, routers establish the SIP session between routers and runs encapsulation for the packets of the mobile terminal. The mobile terminal can maintain communication after a handover because the destination wireless LAN router changes the SIP session. The process for using expanded SIP Mobility method is shown in Figure 4.

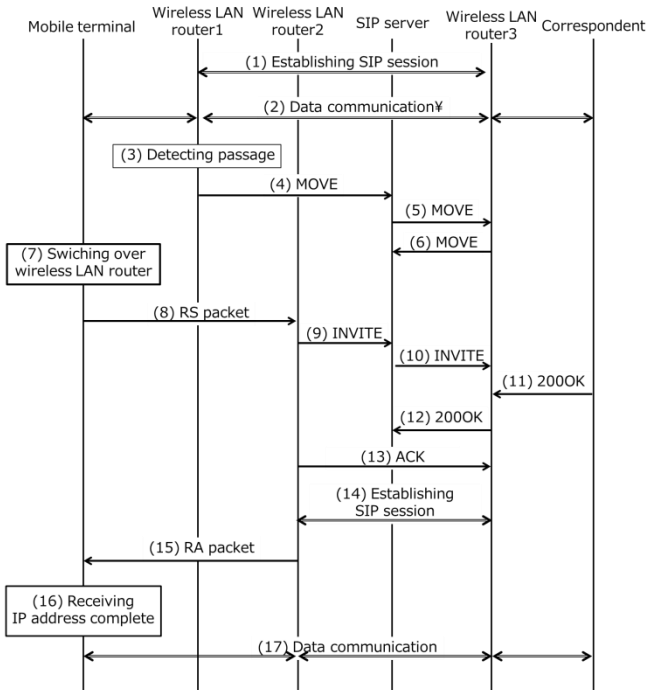


Figure 4: Processing of using expanded SIP Mobility method handover

This process premises on connecting the mobile terminal to the wireless LAN router1, and the mobile terminal is communicating to a correspondent. Wireless LAN routers monitors the RSSI (Received Signal Strength Indication) and it sizes up the movement of the mobile terminal. A SIP session is established between wireless LAN router 1 and wireless LAN router 2, and data communication is run between the mobile terminal and correspondent (Figure 4 (1) - (2)). Wireless LAN router 1 sends the MOVE message to other wireless LAN routers by way of the SIP server when the RSSI drops to a lower value by moving the mobile terminal (Figure 4 (3) - (6)). The MOVE message is the identifier of the mobile terminal and the information. Wireless LAN routers that receive this message keep up a definite period of time. The mobile terminal sends a RA packet to wireless LAN router 2 when it switches over to wireless LAN router 2 (Figure 4 (8)). Wireless LAN router 2, which receives this message, makes a comparison between the identifier of the mobile terminals with the MOVE message and the information. If it agrees with the identifier of the mobile terminals, wireless LAN router 2 establishes the SIP session with the information (Figure 4 (9) - (14)). Wireless LAN router 2 sends a RA packet to the mobile terminal after a definite period of time (Figure 4 (15)). The mobile terminal that receives this packet acquires IP address and proceeds with communication (Figure 4 (16) - (17)).

The SIP message goes through SIP server, but data communication is not needed for the SIP server. Therefore this method uses communication path of choice. In addition, the mobile terminal can use this method without the need for any special features because switching over to the communication path is run between wireless LAN routers. This method accepts a handover in a different network because SIP server exists on the internet.

3.2 Shortening the time spent taking IP addresses

Then we expound shortening the time spent taking IP address. The processing of the current mobile terminal handover is shown in Figure 5.

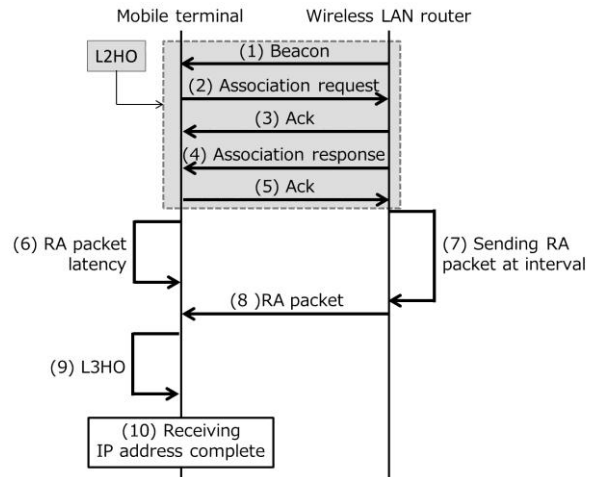


Figure 5: Processing of present mobile terminal handover

The general wireless LAN router broadcasts the Beacon that it passes on the SSID to the mobile terminal (Figure 5 (1)). The mobile terminal that receives the Beacon starts the L2HO and establishes the Association to the wireless LAN router (Figure 5 (2) - (5)). The mobile terminal stands ready to receive the RA packet after the completion of L2HO (Figure 5 (6)). The transmission interval of RA packet is regular, and the setting range of this interval is between 3 and 1800 seconds. The wireless LAN router sends a RA packet to the mobile terminal after a definite period of time, and the mobile terminal runs the L3HO (Layer 3 handover) and gets the IP address (Figure 5 (7) - (10)). If the mobile terminal does not receive the RA packet for a long time, it sends a RS packet to the wireless LAN router 2. Therefore, the handover of the mobile terminal takes over 2seconds after finishing the L2HO because it comes no later than the starting of the L3HO.

The proposal reduces delay that wireless LAN routers send the RA packet to the mobile terminal as soon as the L2HO. The processing of the proposed handover is shown in Figure 6.

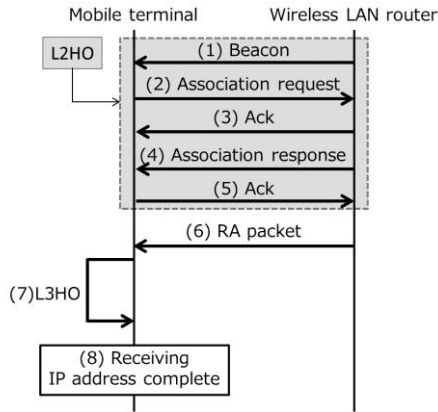


Figure 6: Processing of proposed handover

The L2HO runs conventionally (Figure 6 (1) – (5)). Layer 3 of the wireless LAN router monitors layer 2. The wireless LAN router sends a RA packet to the mobile terminal after receiving the Ack of the Association response (Figure 6 (6)). The delay does not occur between finishing the L2HO and starting L3HO, as the mobile terminal receives the RA packet the L2HO is finished, and the mobile terminal acquires the IP address (Figure 6 (7) – (8)). Therefore, the mobile terminal is shorter than the current mobile terminal, so it proceeds with communication early.

4 EXPERIMENTATION AND EVALUATION

In this chapter, we explain our evaluation. We evaluated the proposed method with the NS2 (Network Simulator version 2) [7] and by running and not running a handover of throughput and the number of packet drops. As a target for comparison is the PMIPv6.

4.1 Comparative evaluation

This section is a comparative evaluation of the PMIPv6 and proposed method. This comparative evaluation is shown in Table 1.

Table 1: Comparative evaluation

	PMIPv6	Proposed method
Bottleneck	Easy to generate	Hard to generate
Running a handover between different networks	Impossible	Possible
Time taking IP address	Long	Short

First, we explain the bottleneck. The PMIPv6 easily converges into a load at the LMA because the mobile terminal runs communication by way of the LMA. The proposed method easily generates a bottleneck because the mobile terminal runs communication without the need for a specific server.

Second, we explain a handover between different networks. The mobile terminal cannot use the PMIPv6 between different networks because it can only run a handover between MAGs connected with the same LMA. The proposed method can establish a SIP session even if the mobile terminal runs a handover between different networks because the SIP server is put on the internet. Therefore the proposed method can be used between different networks.

Finally, we explain the time spent taking IP addresses. PMIPv6 is an amount of time spent taking IP address as in Section 3.2 because it assumes common mobile terminals. The proposed method can take IP address to quickly because wireless LAN routers send a RA packet to the mobile terminal immediately after L2HO by using the mobile terminals. Therefore, the proposed method shortens the time spent taking IP address.

4.2 Experimentation description

In this section, we explain experimentation environment and simulation parameters. We used version ns-2.27 of NS2, and this simulator was run on the Ubuntu-9.04. An evaluation of proposal topology is shown in Figure 7, the evaluation of PMIPv6 topology is shown in Figure 8, and the simulation parameters are shown in Table 2.

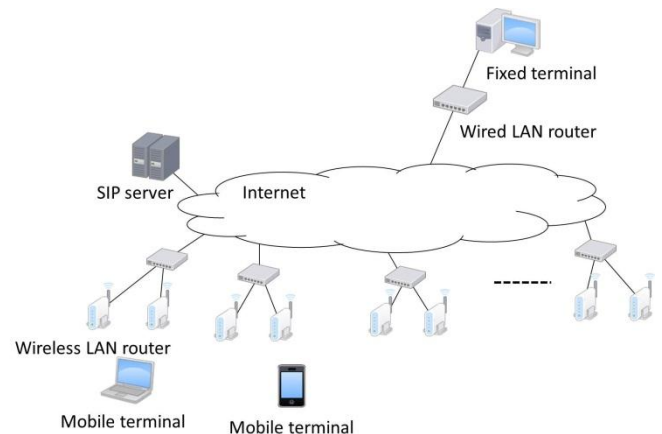


Figure 7: proposed method topology

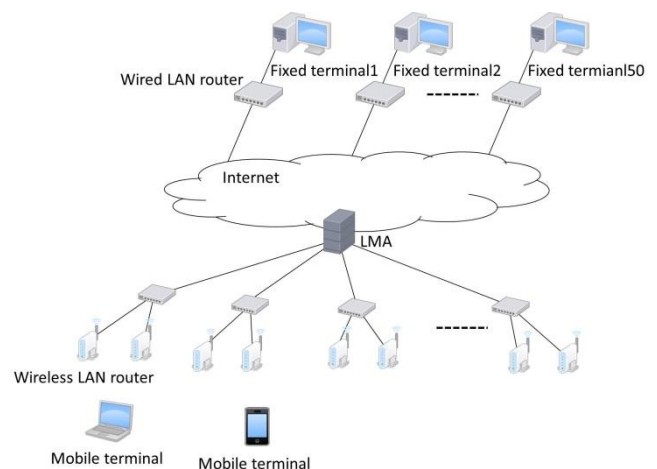


Figure 8: PMIPv6 topology

Table 2: Simulation parameters

Transmission speed of wireless LAN routers (Mbps)	54
Bandwidth (Mbps)	100
Simulation time (sec)	500
Traffic	FTP
Packet size (byte)	1000
Delay the internet (ms)	40
Delay except the internet (ms)	10
Number of wireless LAN routers (routers)	50
Number of fixed terminal (terminals)	50

We referred to references [8] what the topology of PMIPv6 and set the topology of proposed method similar to the topology of PMIPv6.

We assumed a wireless LAN standard of IEEE802.11n, and set the bandwidth that does not become bottleneck. The application protocol is FTP because we assumed the file transfer. We send ping to yahoo.co.jp, determined RTT and set delay the internet with this RTT. Delay except the internet is set lower values than delay the internet because physical distance of except the internet is shorter than the internet.

The communication direction was from a fixed terminal to a mobile terminal. These terminals are one-to-many. The mobile terminal connects the wireless LAN routers. These are one-to-one or one-to-two.

4.3 Evaluation results and examination

This section shows the results and the examination of running and not running a handover. We experimented on several TCP congestion control algorithms. Evaluated TCP congestion control algorithms are Tahoe, New Reno and Vegas. This section shows only Tahoe because there is little difference in these TCP congestion control algorithms.

4.3.1. Not running a handover

The average throughput for 500 seconds is shown in Figure 9, and the number of packet drops is shown in Table 2.

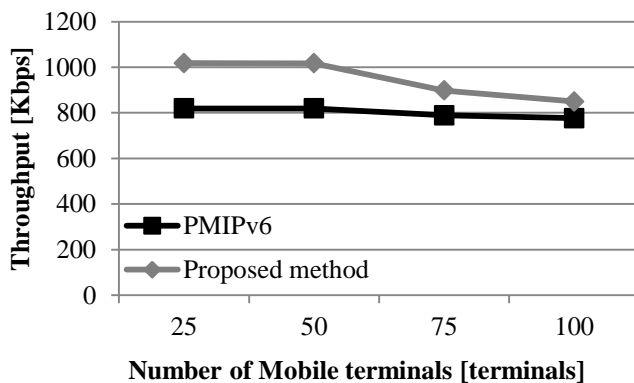


Figure 9: Average throughput for 500 seconds

Table 3: Number of packet drops

Number of Mobile terminal (terminals)	25	50	75	100
Proposal (packets)	16	25	330	913
PMIPv6 (packets)	All	0	0	222
	LMA	0	0	192

For 25 and 50 mobile terminals, the proposed method obtained 1.25 times higher throughput did The PMIPv6. The proposed method did not generate bottlenecks easily because the mobile terminal ran communication without the need for a specific server. However the proposed method had a much higher number of packet drops than did PMIPv6. The proposed method overflowed the queue that connected routers with fixed terminals. This queue overflow is thought to be caused by TCP window size becoming large by communicating without the need for a specific server, and the router's queue reached the maximum number of packets that can be stored. For 75 and 100 mobile terminals, the throughput of the proposed method dropped to a lower value. It is thought that queue overflow occurred often due to the increase in the number of communicating mobile terminals. The queue overflow did not occur due to the one router that connected the fixed terminals. This overflow occurred routers that connected fixed terminals. The most packet drops for PMIPv6 occurred at the LMA. These drops occurred due to converging packets to the LMA. Therefore, it is thought that the communication performance of PMIPv6 deteriorated due to the increase the number of communicating mobile terminals. In the case where a low number of mobile terminals communicated with the same fixed terminal, the communication performance of the proposed method did not get worse but rather improved more than did PMIPv6.

4.3.2. Running a handover

Each mobile terminal ran five handovers in 500 seconds. The average throughput for 500 seconds is shown in Figure 10.

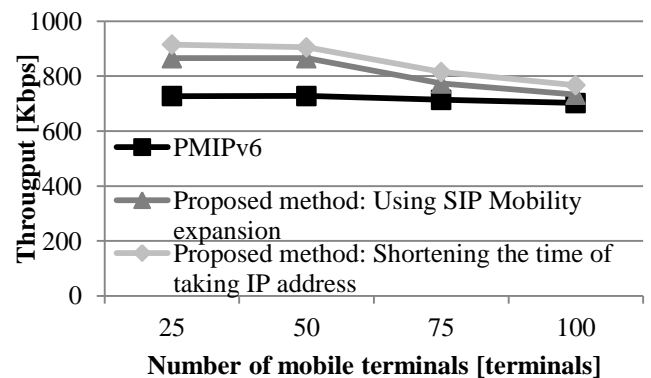


Figure 10: Average throughput for 500 seconds

The proposed method improved the throughput by shortening the time spent taking the IP address because the mobile terminal could run communication quickly after a handover. The throughput of PMIPv6 and the proposed method is nearly equal with the increasing number of mobile

terminals. It can be said that this is said the same as not running a handover. Therefore, the communication performance of the proposed method improved.

5 CONCLUSION

In this study, we purposed a terminal-independent handover method in which the mobile terminal uses a handover technology without the need for any special features and accepted a handover in a different network. We proposed using expanded SIP Mobility method and shortening the time spent taking IP addresses. We experimented with and evaluated the method with a simulator. PMIPv6 impeded on the communication of mobile terminals by going through the LMA. Using expanded SIP Mobility method limited the influence on communication performance by not going through a specific server. The current mobile terminal caused a delay between the completed L2HO and the starting L3HO. Shortening the time spent taking IP addresses cut down on delays by sending RA packet immediately after the completion of the L2HO. We evaluated PMIPv6 and the proposed method. A comparative evaluation showed the effectiveness of the proposed method in that it is rarely generates bottleneck, runs a handover between different networks and shortens the time spent taking IP addresses. Experimentation showed the improvement in the communication performance of the proposed method from using expanded SIP Mobility method and the shortening of the time spent taking IP addresses.

In the future, we will consider evaluating massive topology to show the effectiveness of the proposed method.

REFERENCES

- [1] Internet Society - World IPv6 Day, "<http://www.worldipv6day.org>".
- [2] S. Gundavelli, K. Leung, V. Devarapalli, K. Chowhury, and B. Patil, "Proxy Mobile IPv6," RFC4861, 2007.
- [3] J. Rosenberg, H. Shulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, "SIP : Session Initiation Protocol," RFC3261, 2002.
- [4] H.Miyajima, L.Zhang, H.Hayashi, and T.Fujii, "All-SIP Mobility for Mobile Communications", The Journal of the Institute of Electronics, Information and Communication Engineers, Vol.94, No.1, pp.47-51, 2011. (*in Japanese*)
- [5] Y.Oda, Y.Nakamura, Y.Shiraishi, and O.Takahashi, "A Terminal-Independent Handover Method using SIP Mobility Expansion", Proceedings of the DICO2012, pp.1187-1197, 2012. (*in Japanese*)
- [6] K.Gogo, K.Mitani, R.Shibui, K.Kaneko, J.Ok, S.Komorita, S.Fujimaki, and F.Teraoka, "L3-Driven Fast Handover Using Datalink Layer Information", Information Processing Society of Japan Transactions on MBL, Vol.47, pp.13-18, 2005. (*in Japanese*)
- [7] ns-2, Network Simulator version 2, "<http://www.isi.edu/nsnam/ns/>".
- [8] R.Wakikawa, J.Murai, "The overview of Proxy Mobile IPv6 and its extension", Information Processing Society of Japan Transactions on MBL, Vol.116, pp.83-89, Nov, 2007.

Improvement of The Communication Stability for Wireless M2M Router System

Kazuaki Honda^{*}, and Osamu Takahashi^{**}

^{*}IDY Corporation, Japan
honda@idy-design.com

^{**}Future University Hakodate, Japan
osamu@fun.ac.jp

Abstract -Wireless M2M market is expanding with maturity of a mobile phone infrastructure, and development of the sensor network. Usually, wireless M2M are constituted by one system with an antenna, a radio module, and the router that equipped various interfaces. In order to make it operate as a system without people, the communication stability of a radio module and a router is especially important requirements. In this paper, the radio module and router which are used for this wireless M2M router system were observed. In this paper, the radio module and router which are used for this wireless M2M router system were observed, and requirement for wireless M2M router system was listed and integrated. Moreover, the function mounted in this paper was proved using the system, and was adopted as the settlement-of-accounts machine of coin parking, and the large-scale solar power system.

Keywords: Wireless M2M, Router, Stability, Low cost, Flat-rate, Metered-rate

1 INTRODUCTION

As indicated by the name, wireless M2M refers to “machine-to-machine,” in which communication is carried out between machines using a wireless connection, without going through a human intermediary[1]. The market scale was ¥41.8 billion in FY2010, and grew 13.2% year on year to ¥47.3 billion in FY2011. The market scale is expected to grow an average of 18.4% annually over six years beginning in 2010, reaching ¥115.2 billion in FY2016[2].

As a long tail market, it differs from the high-cost mobile phone market (the traditional customers) in which the average monthly unit price per handset is ¥7,357[3]. It requires a low ARPU with a monthly unit price of a few hundred yen, acquisition cost (customer acquisition cost), and aftercare cost[4]. The potential market for M2M overall, both wired and wireless, is over 20 times the population of Japan, or in other words, a demand of over 3.1 billion[5]. With the reuse of existing products and expansion of the range of application, the market is strong. In addition, in the case of wireless M2M, in which data is sent and received via a wireless connection, the antenna and wireless module and the router that provides each type of interface to control communication and relay transmissions between different networks constitute a single communication control system. This paper defines the aforementioned system as a “wireless M2M router system” (Figure 1, Figure 2). Unlike existing systems that utilize a wired connection, machines connected

to a wireless M2M router system via each type of interface are operated wirelessly at separate locations, without going through a human intermediary. Further, in contrast to consumer services, there is the risk that service will be interrupted if the communication bandwidth is restricted during times of congestion, so the stability of communication and the cost of the system are important.

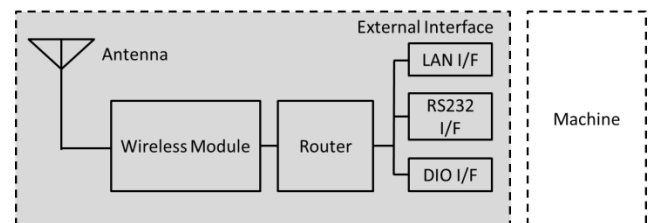


Figure 1 Wireless M2M Router System Block Diagram

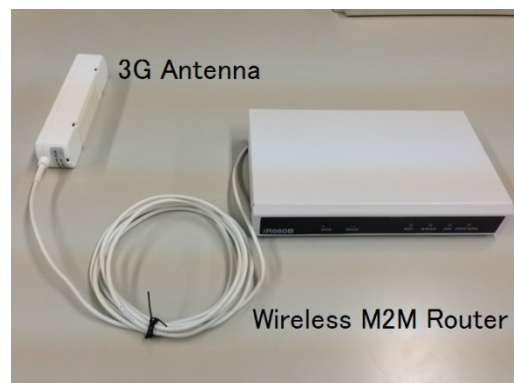


Figure 2 Wireless M2M Router System in the Market

2 PROPOSED METHOD

Reducing the cost of communication is extremely important for the operation of wireless M2M. Recently, flat-rate SIMs with an upper bandwidth limit of 100 Kbps are available for smartphones for a fee of ¥500 or less per month. However, because this type of communication plan is provided as a consumer service, it is not suited to wireless M2M, for which communication cannot be terminated through an upper bandwidth limit or other means during times of communication congestion, etc. Further, best-effort basis flat-rate data contracts provided by individual telecommunications companies have monthly fees that are comparable to those of mobile phones, and the high cost of operation means that it cannot be adopted for this system, which does not go through a human intermediary.

Accordingly, this paper investigates the possibility of use based on a metered packet rate that enable the selection and use of communications costs that are suited to the specific operation, in an effort to provide stable communication services.

Although utilization of a metered packet rate can become a factor in reducing the minimum monthly fee, if the actual number of transmitted packets is high, the cost will ultimately be higher than a flat-rate system, even if the minimum monthly fee is decreased. Because of this, a significant issue will be how the number of packets can be reduced under a metered rate.

Accordingly, in order to consider whether it is possible to build a wireless M2M router system that uses a metered packet rate, we used the wireless M2M router system shown in Table 1 and Table 2 to investigate how many packets not requested by the system are received from outside, other than data that has been sent intentionally by the router program.

Table 1 Wireless M2M Router System Specification 1

Part	Items	Specification
Router	Wireless Module	HUAWEI EM701 HUAWEI EM770J NTT docomo UM-01HW
	Internal I/F	PCI Express mini card
	External I/F	WAN x 1 LAN x 3 RS232 x 1 DIN x 1
	WiFi Type	IEEE802.11n 1T1R
	WiFi Auth.	WEP 64/128bit WPA/WPA2
	Router Function	DHCP Server DMZ IP Sharing MAC Access Control MAX. Connection : 256 Multi BSSID
	Power Supply	5V Φ 2.1 DC Plug
	Size	H 116 x W 189 x H30 (mm)
	Power Consumption	Standby : 4W Operation : 5W (Use 3G/WiFi/WAN/LAN)
	Temp. Humidity	Operation : 0 ~ 50 Degree Storage : -30 ~ 70 Degree Humidity : 10% ~ 95%

Table 2 Wireless M2M Router System Specification 2

Part	Items	Specification
Antenna	Type	Monopole(V), $1/2\lambda$
	Frequency	824MHz ~ 885MHz 1920MHz ~ 2170MHz
	Impedance	50 Ω
	VSWR	< 1.9
	Horizontal Radiation Pattern	Non-directional
	MAX. Gain	824MHz : 1.0dBi 885MHz : -2.1dBi 1920MHz : 0.0dBi 2170MHz : -2.0dBi
	Connector Type	SMA-P
	Cable	2.5m

Withstand Voltage	10W
Operation Temp.	-20 ~ 90 Degree
Water Proof	IPX6

3 INVESTIGATION

Because there are two types of SIM in 3G networks, namely, a SIM that uses a global IP address (global IP) and a SIM that uses a private IP address (private IP), we acquired each set of data for 12 hours each, divided them into well-known ports and non-well-known ports, and compared them. Figure 3 shows the configuration diagram, and Figure 4 shows the actual environment used for the investigation.

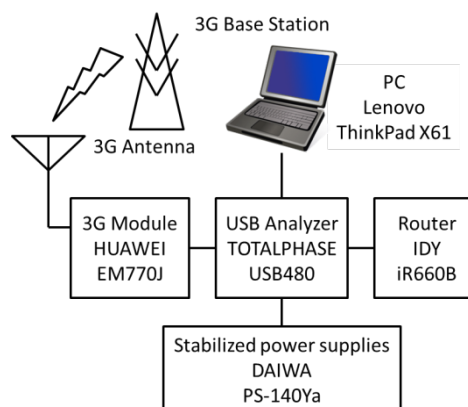


Figure 3 Investigation Environment of Packet Block Diagram



Figure 4 Investigation Environment of Packet

3.1 Investigation of Reception of Unrequested Packets with a Global IP

The following is an investigation of packets not requested by the system that are received from outside. The IP addresses used for the actual confirmation of transmissions will not be disclosed in this paper; however we have indicated the countries to which the IP addresses have been allocated from the IP addresses used to confirm transmissions.

Test date: 2012/06/18 21:39 - 2012/06/19 10:06
Result: Total 100 packets

Table 3 Received packet which is not requested in Global IP.

Kind of Port	Items	Received Count
Well-Known Port	ms-sql-s	8
	ms-wbt-serve	5
	telnet	8
	Sntp	1
	Pmddfing	1
	targus-getdata1	1
	Ssh	2
	z39-50	2
	http	2
	Mcntp	1
	sybase-sqlany	1
	ctp-state	1
	ewdgs	1
	remotedeploy	1
	cisco-sccp	1
	Issd	1
	http-alt	1
	swispol	1
https	1	
Dns	2	
Others	-	13

Table 4 Received packet detail in Global IP.

Sender Country	Total Byte	Frequencies
China	1553	28
United States	437	6
Holland	240	4
Germany	212	4
Turkey	160	3
Taiwan	120	2
South Korea	67	1
Luxembourg	59	1
Saudi Arabia	56	1
Belgium	55	1
Uruguay	55	1
Canada	53	1
Russian Federation	53	1
Hong Kong	53	1

3.2 Investigation of Reception of Unrequested Packets with a Private IP

We investigated the number of packets for which reception cannot be restricted with a private IP, in the same manner as with a global IP.

Test date: 2012/09/06 18:31 - 2012/09/07 13:08

Result: Total 8 packets

Table 5 Received packet which is not requested in Private IP.

Kind of Port	Items	Received Count
Well-Known Port	-	0

Others	-	4
--------	---	---

Table 6 Received packet detail in Private IP.

Sender Country	Total Byte	Frequencies
United States	287	1
South Korea	94	2
Japan	59	1

3.3 Tendencies of Received Packets with Global IPs and Private IPs

As can be seen from the data, 76% of the unrequested packets received by a SIM with a global IP contract during the investigation period were to well-known ports, and the country of transmission varied. However, with a private IP, not a single packet was received by a well-known port, and the only countries of transmission were the United States (Google), Japan (NHN Japan), and South Korea (NHN Corp.). In addition, when using a global IP, the majority of packets sent unilaterally is received by a well-known port, and the possibility that they are malicious packets that target open ports is high. Conversely, with a private IP, the senders were Google or NHN (a LINE service provider), and none of the destinations were well-known ports. Accordingly, because the network used in this investigation is the same as the low-fixed-rate SIMs with limited communication speed that are recently being sold for smartphones, and the applicable SIM cards are assigned private IPs, we can conclude that applications installed from smartphones are randomly sending packets.

3.4 Comparison of the Total Number of Packets with Global IPs and Private IPs

Wireless M2M must be available 24 hours per day, 365 days per year. Accordingly, it is necessary to take into consideration the actual operation, cost, and security when selecting the type of network to use. Calculating the total number of packets for an entire month based on the number of packets for each period of 12 hours used in the investigation, we find that with a global IP there are 5,519 bytes, or approximately 43 packets (calculated with 128 bytes per packet). This becomes 86 packets per day, with 2,850 packets generated over a period of 30 days by access from outside. Similarly, there would be approximately 300 packets over a period of 30 days with a private IP.

The cost benefit of a metered rate over a flat-rate system is generated in the range of 50,000 to 100,000 packets, and based on the observations of this investigation, it appears that the more convenient option can be selected in accordance with the operation specifications as long as well-known ports are closed. The ability to access terminals with a global IP via wireless M2M will likely provide many benefits.

3.5 The Risks of Global IPs and Well-known Ports

Ultimately, there is data received using a global IP without closing well-known ports.

The following shows the number of packets that were sent and received before the session was terminated because the communication session was not closed for a period of 24 hours and data continuously flowed.

When an unrequested packet is sent to a non-listening port, the sent data and received data both have approximately the same number of packets. In the environment used for this packet investigation, only a telnet client, NTP client, and DNS resolver provide the ability to create listen ports. Accordingly, with this data, we can presume that there was some type of access to a listening well-known port telnet, and 7.14 million packets (approximately 872 MB) of data were generated over a single 24-hour session, combining both sent and received data.

Table 7 Illegal Access to Global IP Well-Known Port.

Session Start Time	Session End Time	Transmit Packet	Receive Packet
2012/6/10 1:19:49	2012/6/11 1:19:49	165955.3	6977849.5
2012/6/11 1:21:18	2012/6/11 2:00:30	4328	815.7

4 CONCLUSION

In order to provide stable service in the current wireless M2M market, systems must be built with attention given to the line type and line cost. A stable line is one to which band limitations cannot be applied, and stable line cost requires a communications infrastructure that can be used at a cost that is low enough that it will seem natural for it to be there, similarly to electricity, gas, and water that is continuously received without going through a human intermediary. If, as has been mentioned, the usage fees for a mobile phone have an average monthly unit price of ¥7,357 per month,[3] wireless M2M, which does not go through a human intermediary, will be 1/10 to 1/20 of that amount. The metered rate examined in this investigation is currently the most effective way to meet these requirements for stability. The reason is that the data that is sent and received by machines is, in the first place, sensor information acquired through a machine or commands to control a machine, and by building sent/received packet control that is suited to the system, it is possible to significantly reduce the number of requested packets. Further, through this investigation we were able to determine that it is possible to reduce the reception of unrequested packets. In addition, we were able to obtain a guideline for operation. This will enable the construction of an environment in which small amounts of data will flow naturally like air or water, in a context of an increasing shift to broadband and wireless communication. The utilization of the wireless M2M router system proposed

and verified through this investigation and research has been adopted for a variety of systems, and metered rates are keeping communications costs down. As of July 2013, a cumulative total of 3,000 units have been shipped and are in use. Further, this system is gradually being introduced in coin parking meter systems, home energy management system (HEMS) and building energy management system (BEMS) energy monitoring systems,[6] and electric power confirmation systems in mega solar systems.

ACKNOWLEDGEMENTS

We would like to express our deep appreciation to IDY Corporation's Takeshi Kawatsu, Hiroshi Nakajima, Yosuke Nakayama, Chinatsu Hagiwara, Sayuri Ishikawa, Issei Murakami, and Tomoko Yamada for their cooperation in this investigation and research.

REFERENCES

- [1] ROA Holdings Inc.: "Outlook and Issues for Mobile M2M for B2C: Sensor Networking as a Growth Driver" (2010)
- [2] Business Report Online BRO column: "Rapid Growth to a ¥100 Billion Market Over the Next 5 Years: Researching the 'M2M' Market Which Has Potential" MIC Research Institute Ltd. (November 2012)
- [3] Ministry of Internal Affairs and Communications: "FY2011 Investigation of Differences in Domestic and Foreign Prices Related to Telecommunications Mobile Phones (Comparison of Models)"
- [4] Nokia Siemens Networks Japan Corp.: "Breaking Into the Rapidly Growing Market by Providing Low-cost Services with Standard Support for M2M and Cloud" http://www.nokiasiemensnetworks.co.jp/hirameki/pdf/07_m2m.pdf (March 2013)
- [5] Mobile Computing Promotion Consortium: Mobile M2M Working Group "Announcement of the Launch of Mobile M2M Working Group Activities 1. Background and Goals (Aims) of the Establishment of the Mobile M2M Working Group" (2012)
- [6] The acronyms "HEMS," "BEMS," "CEMS," and "FEMS" are spoken as single words, with the "E" pronounced as in the word "pen." Tokyo Cosmos Electric Co., Ltd. <http://tocos-wireless.com/jp/tech/HEMS.html> (March 2013)



Kazuaki Honda (Regular Member)

Graduated from the Department of Electrical and Electronic Engineering, Kanazawa Institute of Technology in 1995. Joined Uniden Corporation the same year. Developed digital IRD/STB for DirectTV (North America) and SKY PerfecTV! (Japan), etc. Began development of M2M communication devices at PICO Application Corporation in 2001. Began

development of a Windows Mobile dual-mode smartphone at Net-2Com Corp. in collaboration with Fuji Laboratories Ltd. in 2005. Appointed Representative Director and President of Net-2Com Corp. the following year. Currently Representative Director and President of IDY Corporation. Member of the Information Processing Society of Japan.



Osamu Takahashi (Regular Member)

Awarded Master's Degree by the Faculty of Engineering, Graduate School of Engineering, Hokkaido University in 1975. Joined the Yokosuka Research and Development Center of Nippon Telegraph and Telephone Public Corporation (currently NTT) the same year. Engaged in research, development, and standardization of computer networks. Employed at NTT DOCOMO, then in 2004 became a professor at Future University Hakodate. Doctor of Engineering. Received the Achievement Award from the Society. Society fellow. Member of The Institute of Electronics, Information and Communication Engineers and IEEE.

Method to extract Genre or Character from tags of picture via Pixiv

Eiichi Takebuchi[†], Yasuhiro Yamada[†], Akira Hattori[‡] and Haruo Hayami[‡]

[†]Graduate School of Informatics, Kanagawa Institute of Technology, Japan

[‡]Kanagawa Institute of Technology, Japan
{nanashi4129, y.yamada443, akirahattori, hayami.haruo}@gmail.com

Abstract - Any content has tag, which makes a search engine easy to find a particular work. Each tag can be classified into any categories. For example, each of *Dragon Ball* and *One Piece* can be classified as a *genre's name*. Each of *Son Goku* and *Monkey D. Luffy* can be classified as a *character's name*. If a search engine could not distinguish a tag as any category, the search engine and its user cannot recognize a category of the tag when the tag attached to the content is recognized by them.

We researched the method to extract a tag representing genre's name (called a *Genre* after this) and character's name (called a *Character* after this) by analyzing tags of content. The method focuses on inclusion relations between content's tags which are frequently attached to content. To demonstrate validity of the proposed method, we classify Character and Genre according to pictures uploaded to Pixiv¹. Pixiv is one of the famous Social Networking Service in Japan. Its users can upload a fan art to Pixiv. A picture of fan art is involves tags that can be recognized as Character or Genre. And we confirmed that the method enables to classify tags into Character or Genre. This paper states description of the method and an experiment of the method.

Keywords: Folksonomy, Tag, Pixiv, Grouping pictures

1 Introduction

A tags are assigned to the content of a picture, thesis and movie so that a search engine finds a particular works easily. A tag can belongs to any classification. For example, each of *Dragon Ball* and *One Piece* is a genre 's name as the series of work, each of *Son Goku* and *Monkey.D Luffy* is a character 's name.

It will simplify matters, if users can classify a tag on their own as if it is like to attach a tag to any work. In such case, users will be able to classify a tag correctly. Classifications made by users must be useful for a search engine. However, not all of search engines has a faculty of such function.

Developers cannot develop an intelligent search engine in case of that tags are not classified. For example, a system recommends relative works related to particular character of which users would fond by analyzing tags of character 's name. To develop such a system, they will need a classification of Character, otherwise the system cannot recommend any character. Our proposed method extracts genre 's name and character 's name from tags for such a situation.

We researched to classify a tag 's name assigned to any content into a genre 's name or character 's name. Our method

focuses inclusive relations between tags that frequently assigned to works. A fan art represents a character drawn by a creator. Any fan art is based on a particular genre so that a drawn character is a dramatic person on its work.

We verified a validity of a method for to classify tags attached to pictures via Pixiv. The method is able to classify tags into genre's name or character's name. Pixiv is one of the Social Networking Service in Japan. Its users will upload pictures. According to Pixiv, a tag which is frequently attached to the fan art represents genre's name or character's name. We focused the fan art's traits, and researched tags in order to classify them to genre's name and character's name.

Furthermore, we confirmed whether the proposed method classifies genre's name or character's name correctly. According to the verification, we applied the method to a picture which has tag representing a particular genre's name or character's name, and we studied accuracy of the results. Moreover, we studied comprehensiveness of the method by applying it to pictures that were randomly chosen.

The paper describes the method and its experiment.

2 Research background

In this chapter, we defines Genre, Character and the current state of the classification of tags.

2.1 Genre

In this paper, Genre represents a title of the series of works (e.g, Cartoon Animation, Manga, Movie, Novel and so forth). For example, *Harry Potter and the Philosopher's Stone* is not Genre. *Harry Potter* is regarded as a Genre. *The Philosopher 's Stone* is one of the title of the series. However, each of work in the same series may have different outlook on the each work's world. In this case, the Genre represents a work's title. For example, *Precure* is not regarded as a Genre. Both *Heart Catch Precure* and *Smile Precure* are regarded as a Genre of the work's title.

Such concept of Genre was born from *Comic Market* of Doujinshi fair. Doujinshi is a kind of fan art. By all rights, *genre* means *type* or *field*. Booths for selling the goods are divided in order by such category on the site for Comic Market. However, particular fan arts had occupied almost all areas on the site when some fan arts are getting popular. This is why some fan arts had been recognized as an individual Genre for convenience.

¹Pixiv : <http://www.pixiv.net/>

2.2 Character

In this paper, Character represents existence (e.g, human, animal, machine and so forth) that has a proper noun in the work's series. However, the existence who is not related to the stories of the work's series is not regarded as a Character. Accordingly, a human or a car only going through somewhere on the screen is not Character.

Sometimes, a certain tag is regarded as a Genre in the same light as a Character. *Harry Potter* series is representative example. On the one hand, *Harry Potter* is a dramatic person, and on the other hand, it is a Genre also.

2.3 A State of Tag's Classification

Tags was attached to content on user's own. Such tags are utilized to search for content.

A user attaches a tag because the user has any intention to do so. For example, a tag having genre's name or character's name can be classified into a Genre or Character. Accordingly, an existence of a tag must accompany with any reason, and it may be possible to classify such reasons into some superordinate concepts.

Users can attach the tag to the content on the many Web Services. However, it is uncommon that a Web Service permits to classify tags into some categories. Danbooru and Sankaku Channel permit to classify a tag into a Genre and Character. Their users share uploaded pictures. They permit to classify some tags to any classification on the user's own decision.

On the other hand, there are many Web Services whose tags of content are not classified. It is difficult to classify tags newly and immediately for them. They will need to manually or automatically classify tags into some categories. Our target of this research is such the services. The proposed method analyzes the tag's regularities and relations, so that our research conducts automatically extracting tag's classification[1]. This paper describes the method to extract Genres and Characters.

3 Proposed Method

This chapter describes the method to extract classifications of tags of pictures.

3.1 Description

The method is applied to tags of pictures, and extracts Genres or Characters from them. The proposed method have two types, one is the method to extract Genres, the other is to extract Characters.

A common feature of each method is to evaluate a score of each tag. Each method regards its regular result as a tag which has the highest score in whole of results.

According to the method to extract Genres, it focused a tendency for every tag to be attached together each other to particular picture. According to the method to extract Character, it focused a tendency for a tag representing Character attached to a picture together with a Genre.

3.2 Words and functions

This section describes the method of basic words and functions in this paper.

We describe each of tag as t . Tags attached to a picture mean $T = \{t|t_1, t_2, \dots, t_n\}$. n is a number of tags in a picture. In case the pictures is searched (the query is tag), the results are given as $P(x) = \{T|T_1, T_2, \dots, T_m, x \in T\}$. m is a number of the results to search pictures. The number of pictures including the tag is $|t|$.

A tag of Genre is g , and a tag of Character is c . In case such the tag attached to a picture, the picture is described as T_g or T_c .

3.3 The method to extract tags of Genre

The function to extract Genres is $G(T)$. This is given below.

$$G(T) = \max_{t \in T} \frac{\sum_{t' \in T} |P(t') \cap t|}{mn^2} \quad (1)$$

3.4 The method to extract for tags of Character

The method to extract Characters is $C(T)$. This is given below.

$$C(T) = \max_{t \in T} \frac{|P(T_g) \cap t|}{m} \quad (2)$$

Then a tag of Character gotten by $C(T)$, it must satisfies $0 < |T_c| < |T_g|$. If it does not satisfy the condition, we obtain $C(T) = \phi$.

4 Evaluation experiment

We examined accuracy of the method. This experiment evaluates a recall ratio and precision ratio. We set a condition as $n = 10$ and $m = 10$. The reason of $n = 10$, the target of experiment is pictures of Pixiv, and uses of Pixiv is able to attach no more 10 tags to each picture.

4.1 Purpose of experiment

A purpose of the experiment is to assess performance of the method. We assessed its accuracy and comprehensiveness.

Accuracy is an index that represents how correctly tags are extracted from any picture. We evaluated accuracy by applying the method to tags chosen at random, and calculated the ratio that represents how frequently a correct tag is extracted. This verification finds a precision ratio. If precision ratio were high, the method would extract correct tags from any picture.

Comprehensiveness is an index that represents how correctly tags are extracted from pictures which are given arbitrarily. We extracted correct tags beforehand, so that the pictures must include correct tags. If comprehensiveness ratio were high, the method would extract tags from pictures without passing correct tags over.

4.2 Environment of experiment

We constructed the Crawler² for Pixiv as the environment of experiment. This Crawler is implemented by Ruby 1.9.3[6].

Crawler retrieved 1,559,391 existing user ID in ascending order. In addition, Crawler retrieved 1,033,001 user's pictures posted by themselves. And, Crawler retrieved 1,009,544 types of tag, 17,751,104 tags from pictures. We used these datas for the experimnet.

4.3 The method for Genre

This section describes an experiment of the method to extract Genres.

4.3.1 Precision ratio

This paragraph describes an experiment of the method. To evaluate precision ratio, we used a thousand pictures including two types, which given by random sampling. One type indicates that pictures given by random sampling wholly. The other indicates that pictures including Genres given by random sampling. Because, this experiment is used to $m = 1000$. The results are shown below.

Table 1: Precision of the method for Genres

	precision
Random sampling	64.9
To include Genre	56.0

The method is able to extract Genre with a half probability of success as shown in Table 1.

4.3.2 Recall ratio

According to the experiment for evaluation recall ratio, we used three types of Genres. They are *The Melancholy of Haruhi Suzumiya(A)*, *Lucky Star(B)* and *Touhou(C)*. This experiment is used to $m = |t|$. The results are shown below.

Table 2: Recall of the method for Genres

	total	success	recall
A	2733	1,661	0.608
B	4329	3,948	0.912
C	110597	108,938	0.985

Table 2 indicates that the results are separated into Genres with high recall ratio or low recall ratio. As a regularity of Genres with low recall ratio, correct tags were more frequently passed over when there are more pictures containing particular tag than pictures containing the Genre. In particular, there are many tags represents *Original* or *Copyright* in correct tags passed over by the method.

²The Crawler is also referred to as the Bot, Robot or Spider.

4.4 The method to extract for Character

This section describes experiment of the method for Characters.

4.4.1 Precision ratio

We examined three types of Genre to calculate precision ratio. They are *K-ON!(A)*, *Lucky Star(B)* and *A certain scientific railgun³(C)*. The results are shown below.

Table 3: Precision of the method for Characters

	total	success	precision
A	5,341	4,519	0.846
B	4,282	3,561	0.831
C	1,121	997	0.889

The experiment of Table3 indicates that ignored correct tags has many orthographical variant⁴ that were included with an average ratio of 5%. And, 1% of ignored correct tags are attached to a picture which expresses a character not related to the Genre, in spite of the tag of the Genre is attached to the picture. Approximately 5% of ignored correct tags are related to the Coupling⁵.

4.4.2 Recall ratio

In the experiment, we applied the method for Character to pictures. We adopted three types of tag of Characters, they are *Rei Ayanami(A)* (from *Neon Genesis EVANGERION*), *Mikoto Misaka(B)* (from *A certain scientific railgun*) and *Chihaya Kisaragi(C)* (from *THE IDOLMSTER*). The results are shown below.

Table 4: Recall of the method for Characters

	total	success	recall
A	628	628	1.000
B	621	621	1.000
C	1,003	1,003	1.000

We experimented with limited a number of kind of Character. Given the results of Table 4, we confirmed the probability that represents how correctly tags of Character are extracted was 100% when we limited the number of kind of Character analyzed by the method.

5 Discussion

We confirmed that the method for Genres has deficiency in its performance because it was depending on the situation.

³A certain scientific railgun is a common name other than Japan.

⁴Words with the same pronunciation and meaning, but different written forms.

⁵The Coupling is a tag represents a relation between particular two characters. Usually, the Coupling was written in each Character's omitted name in a row. For example, the Coupling of *Mio Akiyama* and *Ritsu Tainaka* is *Mio Ritsu*. First name means the character tend to have activeness. Second name means the character tend to have passiveness.

However, we also confirmed the method for Characters was succeeded because it could extract tags of Characters with high probability.

According to the method to extract Genres, we confirmed a particular regularities between tags leaked from regular results. Expressly, the leakage was caused by tags that were existing more than the number of Genres, which are not related to the targeted Genre. And tags represent *Original* or *Copyright* are also the cause of the leakage. We suppose it might be possible to extend useful filters for the proposed method by considering such the regularities.

The method for Character was the most certain way. A tag related to Coupling summarizes multiple tags. Thus, opinions will be divergent according to whether such tag should be included to a precision ratio or not. Actually, the leakage of tags included many tags of Coupling however. Accordingly, we suppose that it might be possible to grow up accuracy of the method by implementing a filter which recognizes a Coupling.

6 Related works and services

This chapter describes the related works, the relative services and those different points of the proposal method.

6.1 Danbooru

There are many search engine using classified keywords as tags. Danbooru is one of them.

Danbooru is a image board system⁶. It is derived from Futaba Bulletin Board System. There are a lot of clones of Danbooru because it is the open system. Users can attach keywords to pictures on Danbooru. Moreover, it has a wiki. The wiki has many volunteers that explain meaning of each keyword. Users of wiki can assign any category of keyword, Author, Copyright or Character. Danbooru utilizing this classifications which edited on the wiki for the search engine. When tags attached to pictures are classified as Author, Copyright or Character, it can indicate tags with colored state (Figure 1).

Danbooru's classifications of keywords are depending on volunteer's editing. For this reason, it couldn't be the system that has classifications of keywords, If it don't have many volunteers or its users who intend to be volunteers.

The classified keywords are edited manually by users. Our method is able to extract keyword means Genre or Character. Thus, the method doesn't need any manual work to classify keywords into any category.

6.2 Murakami's research

The study reported by Murakami et al[3] shows the way to create a stratified structure from a lot of complicated keywords. The research analyzed a lot of video's tags on Niko Niko Douga⁷. They evaluated relations between a lot of video's tags for their method. Their method utilized ISR(Inter-section Ratio). They described that they can control a precision ratio by using ISR and adjusting the value of threshold(α). On the

⁶It can also be recognized as BBS(Bulletin Board System)

⁷Niko Niko Douga is video hosting service.

Copyrights

? idolmaster 28k
? idolmaster million live! 114

Characters

? nitta minami 116

Artist

? gustav (telomere na) 367

Tags

? 1girl 786k
? bikini 47k
? breasts 317k
? brown eyes 137k
? brown hair 230k
? dutch angle 14k
? front-tie top 4k
? highres 373k
? long hair 513k
? mouth hold 14k
? side-tie bikini 10k
? solo 743k
? swimsuit 72k
? swimsuit pull 1k
? white bikini 1k
? white swimsuit 4k

Information

ID: 1472292
Uploader: dean exia »
Date: 22 minutes ago
Size: 941 KB (848x1200)
Source: pixiv/gustav
Rating: Questionable
Score: 6
Favorites: 5
Status: Active

Figure 1: Danbooru's tags list.

one hand, their method is able to stratify the keywords exactly when the value of α is high that means precision ratio is high. On the other hand, they wasn't able to stratify the keywords appropriately when the value of α is low that means precision ratio is low. Their system set an argument as parent tag. That suggest 10 child tags that have higher ISR value to uses. They succeeded to recommend the unpredictable stratified structure to users by considering conformity noise.

They researched stratification of tags. Stratification of tags is useful because users are able to grasp current directory at which each of them is now. Users are able to understand relations between other tags by looking a parent-child relationship. For example, their method is useful for users, when the users want to know the Characters that appear in the Genre.

However, Murakami's research didn't indicated most adequate value of α . Their prototype system may not suggest stratification adequately because of conformity noise. Their method is not able to extract keywords that can be classified as Genre or Character like our method.

6.3 Wikipedia

We can exemplify the Wikipedia as one of the system related to our study. Wikipedia has many articles with keywords. Wikipedia is made the stratum by setting hyperlinks in the article. Users is understood the position the stratum by wikipedia's stratum.

We had already considered that to classify tags into Genre or Character by using resources Wikipedia has. Actually, lists of the media (Anime, Manga, Movie, ..., etc) are prepared in Wikipedia by users. It may be possible to classify particular keyword as Genre. As another opinion, the method will be able to classify keywords into Character. Wikipedia made the list of Characters in the article that can be recognized as Genre. It may be possible crawling it and get keywords of Character.

These 2 suggestions seems efficient ways. However, they are not able to cope with many orthographical variants in Wikipedia's articles. And, Wikipedia doesn't include each Coupling as its article. Wikipedia is an encyclopedia, for this reason, its users do not add an unpopular and uncommon things or extremely informal things to it as articles. Thus, Wikipedia is an encyclopedia that focuses on defining words correctly, instead of gathering the casual and informal words that users use topically.

Our method extract not only keywords of Genres or Characters, but also orthographical variants, Coupling and topical and informal words. Our method is able to extract words that are more actually used by users than the words Wikipedia has.

6.4 Tag Cloud

Tag cloud is a method to classify keywords extracted from particular contents[4]. It indicates tags of keywords in order to suggest them to users in the order of ascending priorities.

The method extracts keywords as tags from a particular content. This method suggests keywords only. Accordingly, we don't understand that category of each keyword. And the

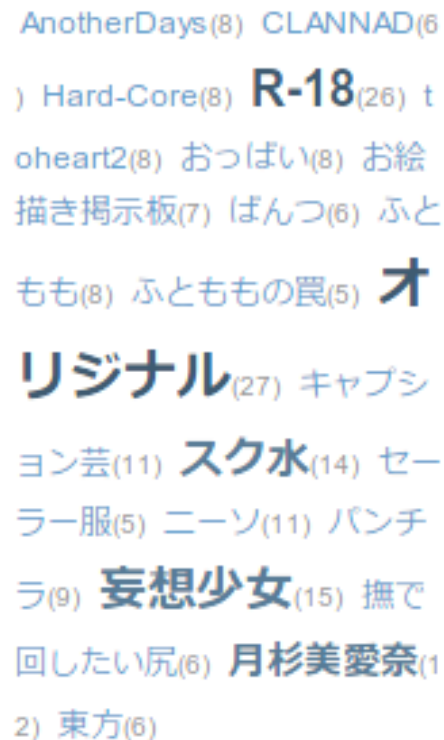


Figure 2: Pixiv's tag cloud by Gisarume (the user number of 266252).

method appends tags to a content automatically in spite of that it is not supposed to be connected to any tag.

On the other hand, our method focuses on classification of keywords. It is able to gather keywords which are supposed to be attached to particular contents and extract keywords that can be classified into Genres or Characters.

6.5 Application of Research

The both method, $G(T)$ and $C(T)$, extract Genre or Character from each set of tags. Each of the method is implemented with high recall ratio. Therefore, the method will be able to extract them with high precision, if the method uses the Machine Learning.

The method is useful for making the dictionary of proper names. In particular, $C(T)$ is able to fit orthographical variants of Character. Therefore, compared with other dictionaries, the method is able to make more accurate orthographical variants. We expect that this dictionary is used by Morphological Analysis, and it would be able to distinguish Character or Genre from contests. And, the machine learning for the image recognition is useful when species data is given. In particular, if the machine learning specializing Genre or Character is implemented, we are able to develop the new image recognition featuring Anime and Comics.

7 Conclusion

This paper described the method to extract Genres or Characters from pictures. A user attaches a tag to a pictures in

accordance with the user's intention. We supposed it is beneficial to improve usefulness of a search engine by proposing a method of classification that considers user's intention.

We noticed that there are regularities between Genres and Characters. A Genre has a tendency to be attached to a picture together with similar tags. A Character also has a tendency to be attached into a picture together with of Genre. Based on these factors, the method focused on relations between many tags, and we researched the method to extract Genres and Characters from pictures.

We implemented experiment for evaluating a recall ratio and precision ratio in each method. According to the method for Genres, we conformed that results of the method depend on the kind of Genre, and confirmed both valid results and invalid results. In contrast, according to the method for Character, we conformed that almost all results were valid results.

We will make a dictionary from Genres and Characters in future research. We aim to build the system which can recommend Genres or Characters by using the dictionary, which considers own user's individuality.

REFERENCES

- [1] Eiichi Takebuchi, Hiroshi Suzuki, Akira Hattori and Haruo Hayami, Pixiv の二次創作イラストに含まれるジャンルタグの自動分類, IPSJ SIG Technical Reports. Groupware and Network services, Vol.86, No.24, pp.1-5 (2013).
- [2] Eiichi Takebuchi, Yasuhiro Yamada, Kazuya Igari, Hiroshi Suzuki, Akira Hattori and Haruo Hayami, Pixiv におけるキャラクタータグを検出する手法の検討, Multimedia, Distributed, Cooperative, and Mobile Symposium 2013. pp.84-88 (2013).
- [3] Naomichi Murakami and Eisuke Ito, Tag hierarchy analysis in Nicovideo movie service, Information Processing Society of Japan. Mathematical Modeling and Problem Solving, Vol.81, No.17, pp.1-6 (2010).
- [4] Yusef Hassan-Montero and Victor Herrero-Solana, Improving Tag-Clouds as Visual Information Retrieval Interfaces, Multidisciplinary Information Sciences and Technologies, InSciT2006 (2006).
- [5] Wikimedia Foundation Inc, "Wikipedia" [Online]. Available: http://en.wikipedia.org/wiki/Main_Page, [Accessed: 31, July, 2013].
- [6] Yukihiro Matsumoto, "Ruby" [Online]. Available: <http://www.ruby-lang.org/en/>, [Accessed: 7, August, 2013].

A Method for Message Delivery in VANET with Scheduled Routes of Vehicles

Masato Nakamura¹⁾, Tomoya Kitani²⁾, Weihua Sun¹⁾,
Naoki Shibata¹⁾, Keiichi Yasumoto¹⁾ and Minoru Ito¹⁾

¹⁾ Graduate School of Information Science, Nara Institute of Science and Technology, Japan

²⁾ Graduate School of Informatics, Shizuoka University, Japan
t-kitani@ieee.org

Abstract - In VANET (vehicular ad hoc network), to deliver messages to a specific location is difficult due to the high mobility of vehicles, and VANET is easy to be disrupted. In this paper, we propose a method for efficient message delivery in VANET utilizing the route information in the car navigation system of each vehicle. In the proposed method, each vehicle periodically exchanges its positional information and scheduled route in the car navigation system with neighboring vehicles in the radio communication range. By referring to the exchanged information, each vehicle forwards a message to the neighboring vehicle that will approach the closest location to the destination of the message. Through simulations about the message delivery ratio, the message overhead and the message delay, we have confirmed that the proposed method achieves better message delivery ratio with less message overhead than the conventional methods: the flooding method, a GeoCast-based method and the Epidemic routing.

Keywords: VANET, message routing, route plan, navigation, DTN

1 Introduction

There are considerable research efforts that made to realize efficient communication between vehicles [1]–[9]. A VANET (vehicular ad hoc network) is a network that uses multi-hop wireless communication among vehicles and devices on roadsides. By using a VANET, various novel services can be realized inexpensively. However, due to the high mobility of the vehicles, it is generally difficult to make a stable data transfer between two specific locations in a VANET. Recently, a technology called a DTN (disruption/delay tolerant network) has been paid attention and it has a potential to improve the message delivery ratio in an unstable network [10]. In a DTN, each node does not forward a message if no suitable relaying neighbor nodes exist in its radio communication range. The node forwards the message later when such suitable nodes appear in its radio communication range [7], [11]. Thus, a DTN can efficiently deliver messages even in the case where the network partition occurs frequently. It can improve the message delivery ratio to utilize the unique characteristics of both a DTN and a VANET, in which nodes can move in high speed while carrying messages and their moving route can be accurately predicted by referring to information in the car navigation system.

In [1], we targeted a vehicular application such that each vehicle collects traffic or weather information and stores the collected data to the roadside device at an intersection. In or-

der to realize such an application, we have proposed a method for delivering messages on a VANET with high message delivery ratio and less communication overhead utilizing the route information in the car navigation system of each vehicle. In the proposed method, vehicles periodically exchange the route information and forward messages to other vehicles that will get closer to the message destination according to the exchanged route information.

In this paper, to evaluate the proposed method, we have conducted simulations using a traffic simulator NETSTREAM [12] and a network simulator to check the performance of the proposed method. We conducted 10 minutes of simulations with a road system of an urban area in a 1.4km × 1.6km region. We have evaluated our method compared with the following three message routing methods: a simple flooding method, Geocast [2] and the epidemic routing [3]. Through the evaluations about the message delivery ratio, the message overhead and the message delivery delay, we have confirmed that the proposed method performs better than the conventional methods.

2 Related Work

A vehicular ad-hoc network (VANET) is a kind of mobile ad-hoc networks (MANET). Especially in VANET, the mobility of nodes is different from that in general MANET, and vehicles run fast and usually go along a road. There are well known routing protocols for MANET including DSR (Dynamic Source Routing) and AODV (Ad hoc On-demand Distance Vector) [4], [13]–[15]. In those protocols, a message is forwarded along a routing path from its source to destination with multi-hop manner, and the path should be connected at a time. In VANET, the connectivity of the network is frequently changing and unstable because of the mobility of the nodes and its heterogeneous density. Thus, the routing protocols for general MANET do not work well in VANET.

DTN (delay/disruption tolerant network) is a network that is easily disrupted, and DTN is used for distributing delay tolerant messages [10]. In DTN, when a node has a message to be delivered and there are no neighboring nodes that are appropriate to relay the message, the node runs to a different place to find another node and to relay the message [3], [5]–[7]. This manner is called *the store and forward* or *the store, carry and forward*. Since vehicles are able to move fast, VANET can adopt the store, carry and forward manner.

As mentioned above, VANET is a DTN. Thus, conventional message delivery protocols for MANET cannot work well in VANET. The most simple message delivery protocol is *the*

flooding method. The flooding method is easy to exhaust the communication bandwidth, and it is a significant problem. *The epidemic routing* protocol[3] is an improvement of the flooding method that lets nodes relay messages with a certain probability, and it reduces the overhead and avoids the congestion of communication.

Each vehicle in VANET can comprehend its location because it is equipped with a car-navigation system and a GPS receiver. GeoCast[2] is a protocol for a vehicle in VANET to select one of neighboring vehicles to relay a message. It considers the geographical distance to the destination of the message. GeoCast reduces the message overhead in the network widely.

The driver in a vehicle equipped with a car-navigation system may go along a route that the navigation system suggested. If neighboring vehicles go along their route plan, we can select relaying vehicles considering the route plan and may reduce the wasted message relaying. GeOpps[8] utilizes this feature.

In GeoCast, a message is relaying along the straight line to its destination. If there are void areas on the line, the message could not be delivered. GeoDTN+NAV[9] is based on GeoCast, and it solves the above problem utilizing the information about the route plan of each vehicle when it meets such void.

Our proposed method utilizes the information about the route plan of each vehicle as same as GeOpps. GeOpps just considers the situation where there are less messages in the network and the congestion does not occur. The proposed method considers the situation where there are lots messages in the network too. GeoDTN+NAV utilizes the route plan just when a vehicle is near the destination of a messages. The proposed method utilizes the route plan more aggressively to improve the message delivery ratio and to reduce the message overhead.

3 Target VANET

In this section, we describe the target VANET model in this paper. A road map is represented as a graph, and a vehicle is represented as a token that transits on the graph. We also describe the assumptions about message delivery and wireless communication.

3.1 Network Model

The target VANET in this paper consists of a road map and vehicles. A road map is represented as a graph $G = (V, E)$ as follows:

V : Set of all intersections. Each element is accompanied with its position.

E : Set of all streets between any two intersections.

The set of vehicles is denoted by C , and each vehicle $c(\in C)$ has the following information:

ID : Vehicle ID.

$route$: Route Plan (A list of $e(\in E)$) that c will go along.

Each vehicle is equipped with the following hardware:

IEEE 802.11 wireless LAN communication device. Its communication radius is denoted by r_c .

Data storage.

Digital road map ($G = (V, E)$).

Route navigation function (can calculate $route$).

GPS receiver.

A message is generated in a vehicle, and it is delivered to its destination. Each message has the important attribute, *deadline*. A message should be delivered to its destination before its deadline. A message has the following information:

ID : ID of the message

b : The destination of message

$deadline$: The deadline to be reached at the destination.

In this paper, we assume that the destination is one of some intersections $B(\subseteq V)$, and $b \in B$. Each of the intersection is equipped with wireless communication infrastructure. Its communication radius is denoted by r_b .

3.2 Message Delivery

As mentioned in Sec. 2, VANET is a DTN, and it adopts the store, carry and forward manner. The vehicles use only one channel during the communication via the WiFi devices. We also assume that each message consists of only one packet.

When a vehicle c enters the communication range, the radius of which is r_c , of another vehicle c' , c and c' can communicate each other. When a vehicle c enters the communication range, the radius of which is r_b , of the roadside unit b in an intersection, c can send a message to b .

4 Proposed Method

In our method, we aim at maximizing the number of messages delivered to the destination before their deadline comes. Here, we assume that a vehicle c can obtain the route plan $c.r$ from its car navigation system, and it strictly follows the route $c.r$. We also assume that the car-navigation system can calculate the shortest distance between a route and the destination of a message. In our method, each vehicle selects appropriate the next vehicle to relay a message considering the route plans of the vehicle and its neighboring vehicles.

The proposed method consists of the following three operations: the information exchanging operation, the message buffering operation and the message forwarding operation, and these operations are executed in parallel.

4.1 Information Exchanging Operation

In this operation, *hello messages* are exchanged between vehicles. A hello message sent from vehicle c consists of the ID $c.ID$, the current position of c and the route information $c.r$. Each vehicle has a table called *neighbor table* to store the received hello messages as shown in Table 1. We store the all contents in a hello message and the time to expire (described below) in a table entry.

Table 1: Example of Neighbor Table

Vehicle ID	Route information	Current pos.	TTL
12	75→31→95→...	(5433,2390)	3
27	75→39→83→...	(5420,2380)	1
55	not equipped with the proposed system	(5430,2385)	-

In this operation, the following two actions are performed every second.

Step 1. Each vehicle broadcasts a hello message every P seconds.

Step 2. We subtract 1 from the “time to expire” of each entry in the neighbor table, and if “time to expire” becomes 0, we delete the entry.

If a vehicle receives a hello message from a neighboring vehicle, the message is stored in the neighbor table.

4.2 Message Buffering Operation

Whenever a vehicle generates a message or receives a message, the vehicle inserts the message into the message queue.

4.3 Message Forwarding Operation

In the message forwarding operation, each vehicle forwards the messages according to the neighbor table. The basic idea of our message routing method is to select the closest neighbor vehicle to approach the destination of a message as relaying vehicle. For example, suppose that vehicle V and W are now located in the top left corner of Fig. 1, and they will follow the route shown in the figure. Since vehicle W will get closer to the message destination than vehicle V , the message will be received and carried by vehicle W .

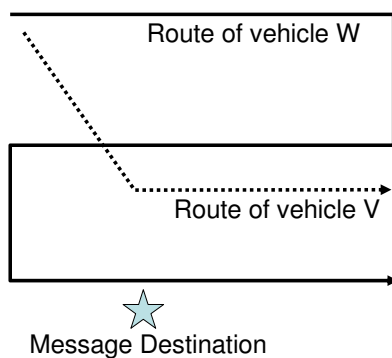


Figure 1: Example of message forwarding

The following process starts when meeting neighboring vehicles.

Step 1. Pop a message from the message queue and check its destination.

Step 2. Calculate the distance between the destination of the message and the closest point on the route of each of neighboring vehicles.

Step 3. Also calculate the estimated time distance to carry the message to the destination with each vehicle.

Step 4. Select the vehicle that can carry the message closest to its destination by its deadline. If the vehicle is itself, keep the message. Otherwise, send the message to the vehicle via wireless communication, and remove the message from the queue.

We assume that the distance can be calculated with a function in the car navigation system.

5 Evaluation

To evaluate the proposed method, we have conducted computer simulation and have evaluated the message delivery ratio and the message overhead, comparing with several conventional methods.

5.1 Simulation Setting

5.1.1 Mobility of Vehicles

To generate the traffic on the road system, we used a traffic simulator NETSTREAM [12] developed by Toyota Central R&D Labs. It has functions for reproducing behaviors of vehicles in a road system taking account of speed limits, intervals of traffic light changes, etc.

The road map for the simulation is shown in Fig. 2. The size of the road map is $1.4\text{km} \times 1.6\text{km}$. All streets in the road system have one lane for each direction. In order to reproduce the situation with disrupt network connectivity and non-uniform traffic density, we generated 462 vehicles in 10 minutes. The maximum number of vehicles in the road map is 166, the vehicle density of which is considerably low.

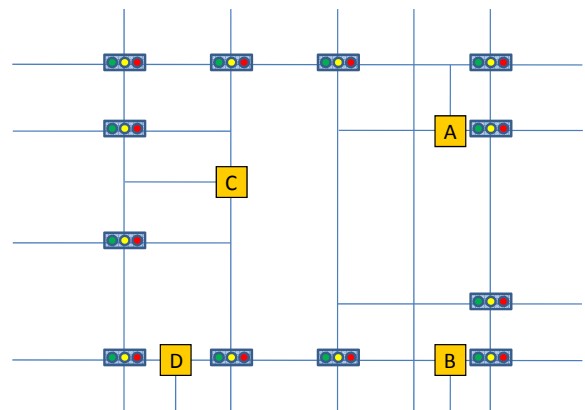


Figure 2: The simulation road map

5.1.2 Wireless Communication

To simulate the wireless network, we developed a simulator which reproduces the behavior of IEEE 802.11 wireless LAN. We used only 1 channel for the wireless communication, and its bandwidth is 12 Mbps.

We assume that both the communication radius r_c and r_b of a vehicle and a roadside unit respectively are 150 meters.

In order to reproduce the behavior of CSMA/CA in the MAC layer, we implemented a carrier detection and collision avoidance mechanism in the simulator. We used the Nakagami fading model [16] to determine if a reception of each packet is successful. In this model, the success rate is determined by the distance between the transmitter and the receiver as shown in Fig. 3.

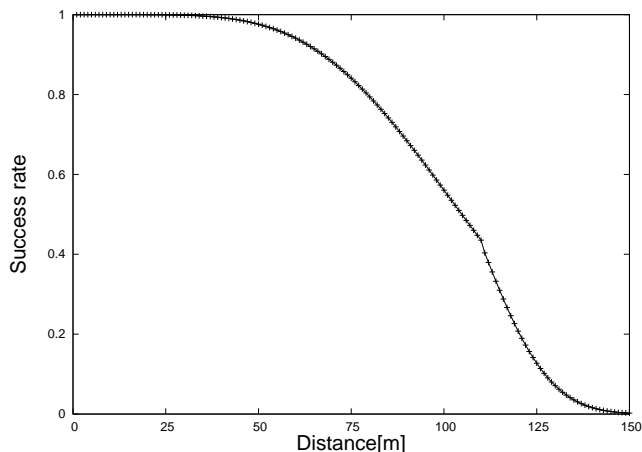


Figure 3: Nakagami fading model

5.1.3 Message

A vehicle generate new message with the rate g . Its unit is [messages/sec/vehicle]. We use the range $[0.1, 0.5]$ for g .

We assume that the deadline of each message is 300 seconds after the message generated. Messages should be delivered within 300 seconds; otherwise the messages will be discarded. Each message is randomly set a destination which is either A, B, C or D in Fig. 2.

The size of a message is 1.5k bytes.

The size of the message buffer (queue) is unlimited.

5.1.4 Route Plan Sharing

We set the parameter P to 3 seconds to broadcast the hello messages because we need to make vehicles exchange messages when they pass each other. If the vehicles move in 16.67m/s (60km/h) and two vehicles pass each other at this speed, they stay within 150m of the radio communication range of each other for 4.5 sec. Thus, we set P to 3 with a 1.5 sec. of safety margin. The size of a hello message is 100 bytes

The all parameters' value are summarized in Table 2.

5.2 Compared Methods

We compared our method with a simple flooding method, the Geocast and the epidemic routing protocols.

5.2.1 Flooding

Whenever a vehicle generates or receives a message, the vehicle broadcasts the message to its neighboring vehicles.

Table 2: Simulation parameters

Area size	1.4km \times 1.6km
Max speed of vehicles	60km/h
Number of crossings	26
Number of vehicles	462
Number of the messages	500 – 5000
Radio range	150m
WiFi bandwidth	12Mbps
Hello message size	100byte
Message size	1.5kbyte
The time duration to the deadline	300 seconds
Buffer (queue) size	unlimited
Simulation period	10 minutes

In the case that the number of vehicles and the message generation ratio are low and there are no message congestions, in other words there are no limit of bandwidth and the size of message buffer, the flooding method performs the upper bound of the message delivery ratio. In most cases, the flooding causes message congestion and it degrades the performance.

5.2.2 GeoCast

In GeoCast, a vehicle exchanges the location information of neighboring vehicles by hello messages. Each vehicle relays a message to the geographically closest vehicle to the destination of the messages. GeoCast can reduce the message overhead significantly.

5.2.3 Epidemic Routing

Epidemic routing is like a probabilistic flooding. To select the appropriate rate, the epidemic routing works well.

In this paper, we use 30% as the relaying probability.

5.3 Evaluation Items

To evaluate the methods, we have derived the following metrics: the message delivery ratio, the message overhead and the average message delivery delay.

The message delivery ratio is calculated as the number of delivered messages by their deadline divided by the number of generated messages.

The message overhead is calculated as the sum of the total amount of hello messages and the total amount of relayed messages.

The average message delivery delay is calculated from only the delivered messages to their destination.

5.4 Simulation Results

We have conducted the following three types of simulations: Changing the message generation ratio, changing the penetration ratio of the system and changing the size of the message buffer (queue).

5.4.1 Changing Number of Messages

First, we evaluated the methods changing the message generation ratio.

Figure 4 shows the message delivery ratio. The proposed method performs the best. When the message generation ratio is low, the flooding also performs as well as the proposed method, but the message generation ratio gets higher, the message delivery ratio of the flooding degrades more because of the message congestion.

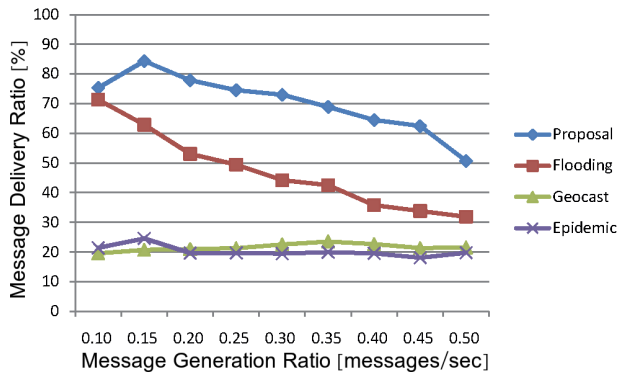


Figure 4: Messages generation ratio vs. Message delivery ratio

Figure 5 shows the message overhead. The message overhead of the flooding is much larger than the other methods, so the flooding is easy to cause the message congestion. The message overhead of the proposed method is as low as Geo-Cast.

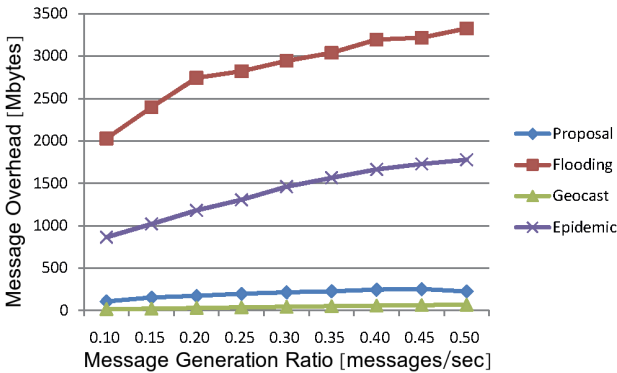


Figure 5: Messages generation ratio vs. Message overhead

Figures 6 and 7 show the average message delivery delay and the commutative message delivery ratio, respectively. As shown in the figures, the proposed method utilizes the delay tolerance (the time margin to the deadline) in order to improve the message delivery ratio and to reduce the message overhead.

5.4.2 Changing System Penetration Ratio

Seconds, we evaluated the methods changing the penetration ratio of the proposed method. It is difficult to disseminate the

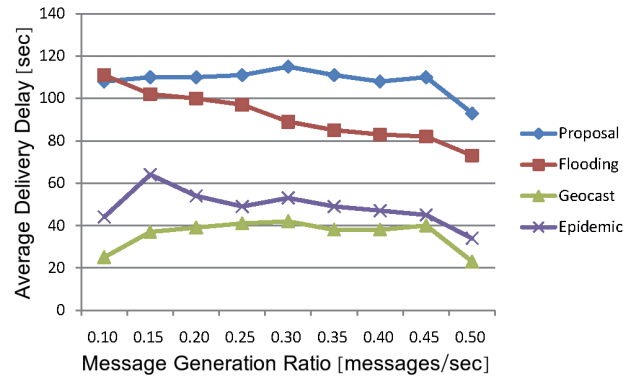


Figure 6: Messages generation ratio vs. Average message delivery delay

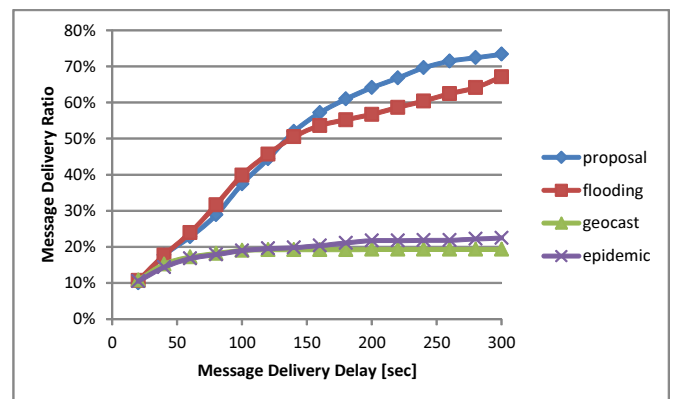


Figure 7: Message delivery delay vs. Commutative message delivery ratio

system to all vehicles. To make the proposed system practical, the system should work well in the situation that the system and other conventional systems are mixed. Here, we assume that a vehicle is equipped with the epidemic routing method if the vehicle is not equipped with the proposed method. The message generation ratio is 0.10 messages/sec./vehicle.

Figure 8 shows the message delivery ratio. The penetration ratio gets higher, the proposed method performs better.

Figure 9 shows the message overhead. When the penetration ratio is less than 70%, the message overhead of the proposed method is higher than the epidemic routing. However, the penetration ratio is more than 70%, the proposed method reduces the message overhead widely, keeping the message delivery ratio high.

Figure 10 shows the average message delivery delay.

5.4.3 Changing Storage Size

Lastly, we evaluated the methods changing the size of the message buffer (queue). Heretofore, the size of the message buffer was unlimited. Here, the size is limited, and if a message is generated or received and the message buffer is full, the oldest message is discarded and the new message is put into the message buffer. The data generation ratio is 0.10 messages/sec./vehicle.

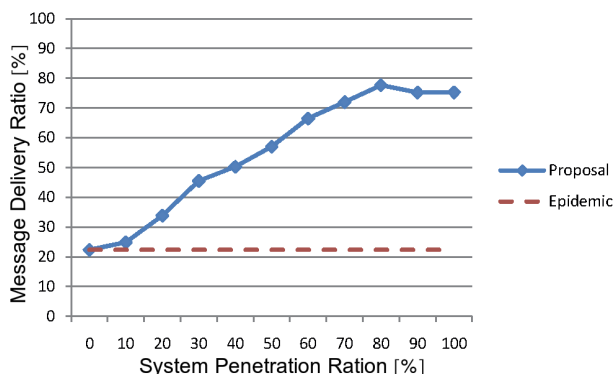


Figure 8: Penetration ratio vs. Message delivery ratio

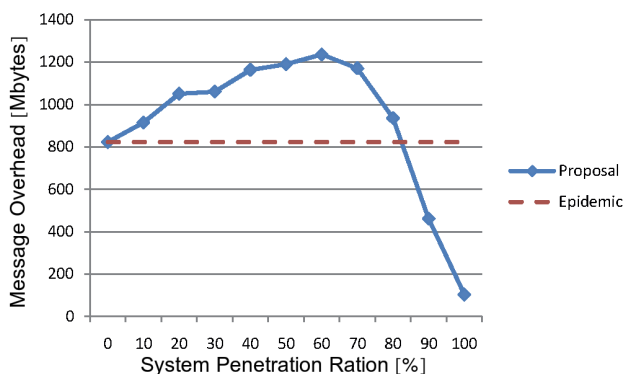


Figure 9: Penetration ratio vs. Message overhead

Figure 11 shows the message delivery ratio. The proposed method performs the best. It is interesting to note that the flooding method gets better when the size of the message buffer gets larger. It is because the flooding method broadcasts lots of messages and the vehicles need to keep such amount of messages.

When the message generation ratio is low, the flooding also performs as well as the proposed method, but the message generation ratio gets higher, the message delivery ratio of the flooding degrades more because of the message congestion.

Figures 12, 13 and 14 show the message overhead, the average message delivery delay and the commutative message delivery ratio, respectively.

6 Conclusion

In this paper, we proposed a DTN-based method for delivering messages utilizing the route information in car navigation systems. Our simulation-based evaluation showed that our method achieves a higher message delivery ratio with less message overhead compared with some existing routing protocols.

As a future study, we will design and implement a protocol to deliver a message from a road side device to a specific vehicle.

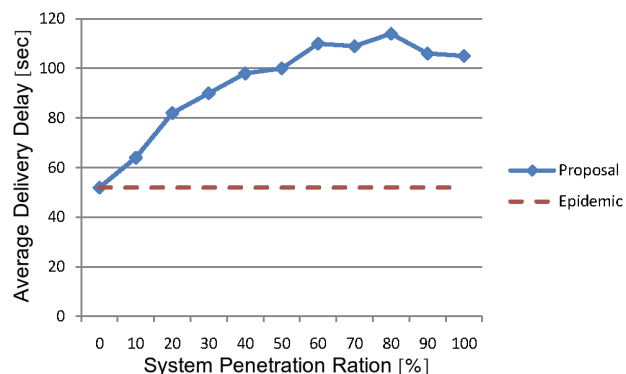


Figure 10: Penetration ratio vs. Average message delivery delay

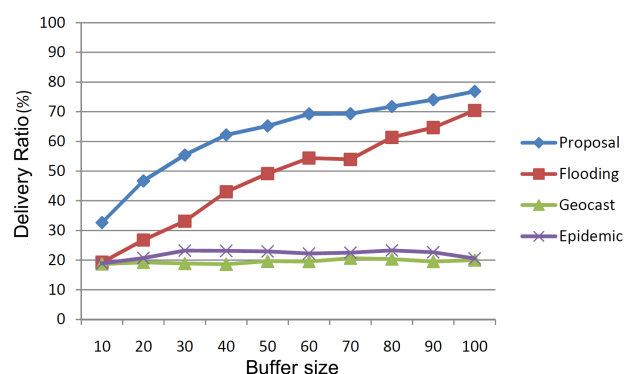


Figure 11: Size of message buffer vs. Message delivery ratio

REFERENCES

- [1] M. Nakamura, T. Kitani, W. Sun, N. Shibata, K. Yasumoto and M. Ito, "A Method for Improving Data Delivery Efficiency in Delay Tolerant Vanet with Scheduled Routes of Cars," *Proc. of the 7th IEEE Consumer Communications and Networking Conference (CCNC2010)*, pp. 1009–1013, 2010.
- [2] Y.B. Ko and N.H. Vaidya, "Geocasting in mobile ad hoc networks: location-based multicast algorithms," *Proc. of IEEE Workshop on Mobile Computing Systems and Applications (WMACSA'99)*, pp. 101–110, 1999.
- [3] A. Vahdat and D. Becker, "Epidemic routing for partially-connected ad hoc networks," *Duke University Technical Report CS-2000-06*, 2000.
- [4] W. Sun, H. Yamaguchi, K. Yukimasa and S. Kusumoto: "GVGrid: A QoS Routing Protocol for Vehicular Ad Hoc Networks," *Proc. of 14th IEEE International Workshop on Quality of Service*, pp. 130-139, 2006.
- [5] N. Shibata, T. Terauchi, T. Kitani, K. Yasumoto, M. Ito and T. Higashino: "A Method for Sharing Traffic Jam Information Using Inter-Vehicle Communication," *Proc. of International Workshop on Vehicle-to-Vehicle Communications (V2VCOM) (MobiQuitous2006 Workshop)*, (CD-ROM), 2006.
- [6] T. Kitani, T. Shinkawa, N. Shibata, K. Yasumoto, M. Ito and T. Higashino: "Efficient VANET-based Traffic In-

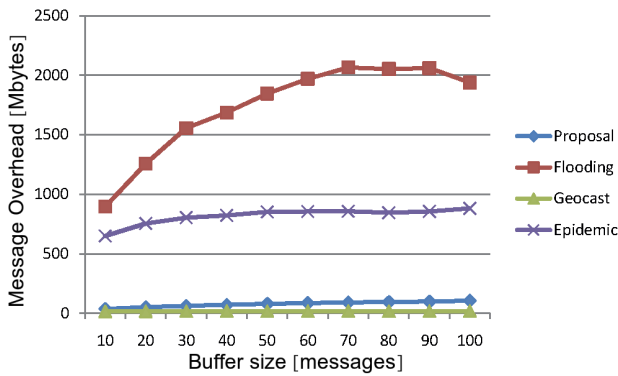


Figure 12: Size of message buffer vs. Message overhead

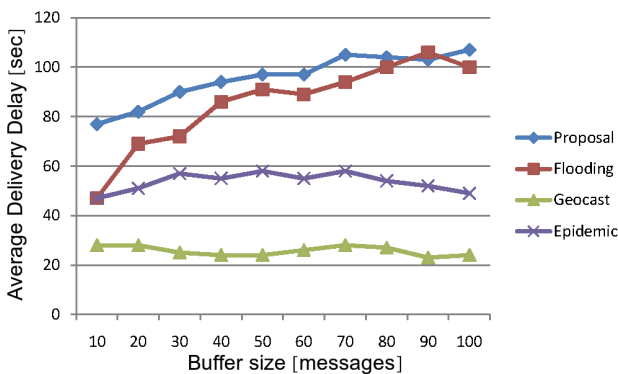


Figure 13: Size of message buffer vs. Average message delivery delay

formation Sharing using Buses on Regular Routes,” *Proc. of 2008 IEEE 67th Vehicular Technology Conference (VTC2008-Spring)*, pp. 3031–3036, 2008.

- [7] T. Spyropoulos, K. Psounis and C.S. Raghavendra: “Spray and Wait: An Efficient Routing Scheme for Intermittently Connected Mobile Networks,” *Proc. of ACM SIGCOMM workshop on Delay-tolerant networking (WDTN-05)*, pp. 252–259, 2005.
- [8] I. Leontiadis and C. Mascolo: “GeOpps: Geographical Opportunistic Routing for Vehicular Networks,” *Proc. of IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM 2007)*, pp. 1–6, 2007.
- [9] P.-C. Cheng, K.C. Lee, M. Gerla and J. Härrri: “GeoDTN+Nav: Geographic DTN Routing with Navigator Prediction for Urban Vehicular Environments,” *Journal of Mobile Networks and Applications*, Vol. 15, Issue 1, pp. 61–82, 2010.
- [10] Delay Tolerant Networking Research Group, <http://www.dtnrg.org/wiki/>.
- [11] J. Sushant, F. Kevin and P. Rabin: “Routing in a delay tolerant network,” *Proc. of ACM SIGCOMM Computer Communication Review*, pp. 145–158, 2004.
- [12] E. Teramoto, M. Baba, H. Mori, H. Kitaoka, I. Tanahashi and Y. Nishimura: “Prediction of Traffic Conditions for the Nagano Olympic Winter Games Using Traffic Simulator: NETSTREAM,” *Proc. of 5th World Congress on Intelligent Transport Systems*, Vol. 4, pp. 1801–1806, 1998.

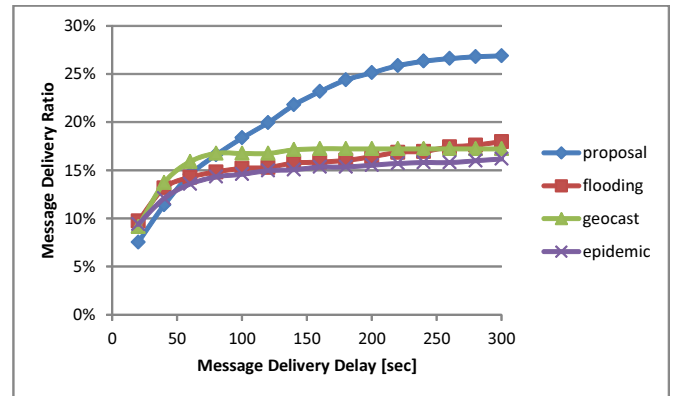


Figure 14: Size of message buffer vs. Commutative message delivery ratio

Traffic Simulator: NETSTREAM,” *Proc. of 5th World Congress on Intelligent Transport Systems*, Vol. 4, pp. 1801–1806, 1998.

- [13] D. Johnson and D. Maltz: “Dynamic Source Routing in Ad Hoc Wireless Networks,” *Mobile Computing, Kluwer*, pp. 153–181, 1996.
- [14] C.E. Perkins and E.M. Royer: “Ad hoc on-demand distance vector routing,” *Proc. of 2nd IEEE Workshop on Mobile Computing Systems and Applications (WMACSA’99)*, pp. 90–100, 1999.
- [15] C.E. Perkins and P. Bhagwat: “Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers,” *ACM SIGCOMM Computer Communication Review*, Vol. 24, Issue 4, pp. 234–244, 1994.
- [16] M. Killat, F. Schmidt-Eisenlohr, H. Hartenstein, C. Ressel, P. Vortish, S. Assenmacher and F. Busch: “Enabling Efficient and Accurate Large-Scale Simulations of VANETs for Vehicular Traffic Management,” *Proc. of the 4th ACM International workshop on Vehicular ad hoc networks (VANET2007)*, pp. 29–38, 2007.

Panel Discussion

IWIN2013 Panel Session

September 3, 2013
Yoshimi Teshigawara
Tokyo Denki University

SOCIAL COMPUTING

Panelists

- Prof. Sadao Obana
The University of Electro-Communications
- Prof. Teruo Higashino
Osaka University
- Dr. Hiroshi Inamura
Research Laboratories, NTT DOCOMO Inc.

Discussion points

- Which area are you involved in?
- What issues should be considered?
- How they will be solved, if possible?
- What are the future prospects?
- Others

What is Social Computing?

- Social computing** is a general term for an area of **computer science** that is concerned with the intersection of **social behavior** and **computational systems**. It has become an important concept for use in business.
- Social computing refers to the use of "Social software"
- Enables people to connect or collaborate through computer-mediated communication and to form online communities.

Definition of Social Computing

- "Social Computing" refers to systems that support the gathering, representation, processing, use, and dissemination of information that is distributed across social collectivities such as teams, communities, organizations, and markets.
- Moreover, the information is not "anonymous" but is significant precisely because it is linked to people, who are in turn linked to other people.¹

http://en.wikipedia.org/wiki/Social_computing

What is Social Computing?

The diagram illustrates various social computing services categorized into several groups:

- Social Networking:** LinkedIn, Facebook, MySpace, del.icio.us, Digg, StumbleUpon
- Blogging & Micro Blogging:** Twitter, Socialtext, Yammer, Jive, Huddle, Basecamp
- Video & Music Streaming:** YouTube, iPlayer, Last.fm, Flickr, Photo & Presentation Sharing, Picasa, Slideshare
- Wikis:** Wikipedia, WikiHow, Google, Bing, Yahoo!
- Email and Instant Messaging:** Gmail, Hotmail, Youko, Mail, FeedBurner, NewsGator
- Ideas & Innovation:** Spigit, IdeaStorm, UserVoice, Webex, Live Meeting
- Social Bookmarking:** del.icio.us, Digg, StumbleUpon
- Forums & Communities:** Jive, Huddle, Basecamp
- Photo & Presentation Sharing:** Flickr, Picasa, Slideshare
- Web & Enterprise Search:** Google, Bing, Yahoo!
- RSS & News Syndication:** FeedBurner, NewsGator
- Online Meetings & Telephony:** Webex, Live Meeting

http://www.deloitte.com/assets/Dcom-Mexico/aca/RelatedAssets/Documents/mx/tes/mxSocial_computing.pdf

Examples

- Online Identities
social identity that network users establish in online communities
- Blogs
- Social Networking Websites
- Social Media Sharing
- Social Network Analysis
collecting data from social network websites
Use of semantic web technologies to determine social behavioural patterns

<http://www.authorstream.com/Presentation/hanuman108-1698944-social-computing/>

7

In addition

- Collaborative e-learning
- Face-to-face meeting
play a big role in social computing

8



9

From M2M to H2H

- Networks accelerate connections or associations among people
- H2H
Human to Human
Internet of People (IoP)
- Collaborative work
Hand to Hand
- Establish emotional bonds
Heart to Heart

10

Issues to be considered

- Trust and Security
- Privacy
- Intellectual property
- Energy –saving
- Human-centered
- Unselfconsciously supported

11

Requirements for Social Computing

- Usability
- Comfortableness
- Use and Gratification, satisfaction

In addition,

- Pleasantness
- Euphoria, feeling of happiness

12

Prospects for Social Computing

- Provision of more realistic face-to-face communications
- 3D Internet
- Sensible networks to five senses touch, taste, hearing, eyesight, smell
- Possible to shake hands and hug
- In the future, sensing of six sense

13

Utilization of 3D printers

- Real objects will be presented.
- Authentic-looking faces and hands are used.

14

THANK YOU FOR YOUR KIND ATTENTION.



Panel

Social Computing

Teruo Higashino
Osaka University

+ Spread of Social Computing

- Social Computing: the intersection of social behavior and computational systems
 - weaker sense (tools for supporting social behavior): blogs, email, instant messaging, social network services, wikis, social bookmarking, cookpad
 - stronger sense (wisdom of crowds): collaborative filtering, online auctions, prediction markets, reputation systems, computational social choice, tagging, and verification games
- James Surowiecki (2004)
 - The Wisdom of Crowds (“集合知” in Japanese): Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations

+ Usage of Social Computing

- In order to carry out collaborative filtering and computational social choice, we need to collect several knowledge from many people. Then, such knowledge is got in shape as “the wisdom of crowds”.
 - e.g.: tripadvisor, tabe-log, cookpad
- Crowd computing (from Wikipedia)
 - Crowd computing is an overarching term which defines the myriad human interaction tools that enable idea sharing, non-hierarchical decision making and the full utilization of the world’s mind space. Examples of these tools (many falling under the Web2.0 umbrella) include collaboration packages, information sharing software, such as Microsoft’s SharePoint, wikis, blogs, alerting systems, social networks, SMS, MMS, Twitter, Flickr, and even mashups. Business and society in general increasingly rely on the combined intelligence, knowledge, and life experiences of the “crowd” to improve processes, make decisions, identify solutions to complex problems and monitor changes in consumer taste.


+ How we can construct the wisdom of crowds

- Important factors:
 - Good tools for collecting valuable knowledge
 - Incentive for individuals so that they can provide valuable information
 - Intellectual filtering mechanisms for removing wrong/invaluable information
 - Democratic editing/mining/decision mechanisms
 - etc.
- It is very important for individuals to know real situations for their target, which include not only good aspects but also bad ones.

+ Research for Social Computing

- Following research themes are considered:
 - Development of software tools for collecting valuable knowledge
 - e.g.: lecture support tool: Teacher and his/her students can jointly make the wisdom of crowds for the lecture so that better lecture can be provided
 - Development of techniques for data mining from a large amount of data
 - e.g.: Big voice should not mean big influence. Intellectual and democratic mining mechanisms are needed.
 - Development of the wisdom of crowds using pervasive sensing from the real world
 - e.g.: more intellectual driving assistance, guidance for travellers, evacuation from disasters
 - Usage of multiple social computing information
 - e.g.: cookpad is useful for making dinner tonight. If we can combine it with information from neighboring supermarkets, one might make nice dinner with low costs.

+ Thank you very much



Social Computing for Collaboration

Hiroshi Inamura
NTT DOCOMO
Research Labs.

What is Social Computing?

- Connecting people, maintain relationship – SNS?
 - Deriving useful information from people for the people
 - Recommendation, Preference in Similar and Closer people
- Next step?
 - What we can **do with** the social computing?
 - May be a social group with **productive goal**?
 - Do more with “intersection of **social behavior and computational system**”

Kasparov v.s. world

- Challenge: Beat the world champ with large number of less ranked players in chess game
 - **Anyone in the world can participate** to vote the next move
 - 75 country, 50,000 people, 5,000 vote per move
 - Fund by Microsoft and First USA bank, 1999
 - Contrary to expectations, the game produced a mixture of **deep tactical and strategic ideas**, and although Kasparov won, he admitted that he had never expended as much effort on any other game in his life
- Reasons for the Success
 - Taking enough time for collaboration between game (24H)
 - Organize information for new participants in communication forum.
Recruit point experts
 - With rough consensus and contribution, a leading person emerged to **steer discussion**
 - A women US chess champ, Iriana Krush, 15yr old at the time

Social Computing for Open Collaboration

- Vision:
 - The social computing for **organizing mass-human capability to focus on specific goal.**
- Goal:
 - Effective collaboration
 - win a game, problem solving...
 - Wider view of the world
 - participatory sensing, defining problem in society...
- How:
 - Mass-collaboration
- What is **Productive Goal**?

More data? Open Data Innovation

- Open Data, Open Government
 - The computing for assisting the collaboration of participants
 - Collect **ideas for problem definition for the society**
 - In US
 - Data.gov, Code for America
 - In Japan,
 - 「オープンデータ流通推進コンソーシアム」
～公共情報等のオープンな利活用の実現により社会・地域・産業の活性化を推進～
- With social computing,
 - Collect data for **surrounding environment**
 - Traffic Jam, Graffiti, etc...

Conclusion

- We can define, solve problem in society with the **power of crowds**
 - Social Computing for Open Collaboration, Open data
 - **Open data X Collective Intelligence X Collective Sensing**
- ICT community can contribute here with
 - **Novel computing tools**
 - to encourage mass-collaboration
 - to analysis data to visualize the changes in issues
 - to collect data surrounding environment
- **Code for Japan ?**