2008

Proceedings of
International Workshop on Infomatics

September 9-11, 2008
Wien, Austria

Sponsored by Informatics Society

2008

Proceedings of
International Workshop on Informatics

September 9-11, 2008
Wien, Austria

Sponsored by Informatics Society

URL: http://www.infsoc.org/

Publication office
Informatics Laboratory
3-41, Tsujimachi, Kitaku, Nagoya 462-0032, Japan
Publisher
Tadanori Mizuno, President of Informatics Society
ISBN:
978-4-902523-18-8
Printed in Japan

Table of Contents

Session 4: Design Methodology (Chair Jun SAWAMOTO) (Sep.10 13:20-15:20)

Session 5: Groupware (Chair Kenichi OKADA) (Sep.10 15:30-17:00)

Session 6: Social System (Chair Teruhisa ICHIKAWA) (Sep.11 10:00-11:30)

Note: * represents short paper contribution.

A Message from the General Chair

It is my great pleasure to welcome all of you to Wien, Austria for the Second International Workshop on Informatics (IWIN 2008). This workshop is sponsored by the Informatics Society.

The first workshop was successfully held in Napoli, Italy in September 2007, where a variety of interesting papers were presented and eager discussion was made over 10 technical presentations. Similar to the last workshop, many high-quality 27 papers were submitted to the workshop from experts. In order to further raise the quality of the workshop, careful selecting procedure was conducted by the program committee, and only 21 were accepted in this workshop. Based on the 21 selected papers, 6 technical sessions have been organized in a single track format, which highlight the latest results in research areas such as the networking, business system, education system, design methodology, groupware, and social system.

I would like to thank all of participants and contributors who made the workshop possible. It was indeed an honor to work with 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 wish you all a great and enjoyable meeting in Wien, Austria.

Osamu Takahashi

Osamu TAKAHASHI
General Chair
The International Workshop on Informatics

Organizing Committee

General Chair

Osamu Takahashi (Future University-Hakodate, Japan)

Steering Committee

Norio Shiratori (Tohoku University, Japan)
Tadanori Mizuno (Shizuoka University, Japan)
Teruo Higashino (Osaka University, Japan)
Jun Munemori (Wakayama University, Japan)
Toru Hasegawa (KDDI R&D Laboratories, Japan)
Ken-ichi Okada (Keio University, Japan)

Program Committee

Kouji Yoshida (Shonan Institute of Technology, Japan)
Gen Kitagata (Tohoku University, Japan)
Hiroshi Mineno (Shizuoka University, Japan)
Hirozumi Yamaguchi (Osaka University, Japan)
Wang Xin (Fudan University, China)
Teruyuki Hasegawa (KDDI R&D Laboratories, Japan)
Chakraborty Debasish (Tohoku University, Japan)
Tomoya Takenaka (Shizuoka University, Japan)
Kazuya Matsudaira (Shizuoka University, Japan)
Teruhisa Ichikawa (Shizuoka University, Japan)
Kunihiro Yamada (Tokai University, Japan)
Yuko Murayama (Iwate Prefectural University, Japan)
Jun Sawamoto (Iwate Prefectural University, Japan)
Hisao Koizumi (Tokyo Denki University, Japan)
Yohtaro Miyanishi (Miyagi University, Japan)
Masahiko Ishino (Mitsubishi Electric Information Technology Corp., Japan)
Yoshia Saito (Iwate Prefectural University, Japan)
Nobutsugu Fujino (Fujitsu Laboratories, Japan)
Fumiaki Sato (Toho University, Japan)

Session 1: Networking
(Chair Nobuhiro KATAOKA)

On harmonizing IMS with MPLS-based traffic engineering

Takeshi Usui, Yoshinori Kitatsuji, Teruyuki Hasegawa, and Hidetoshi Yokota

KDDI Laboratories, Inc. 2-1-15 Ohara, Fujimino City, Saitama Prefecture 356-8502 JAPAN
ta-usui@kddilabs.jp

Abstract -IMS is a key technology for next generation network, which enables NSPs (network service providers) to provide various services over IP-based fixed and mobile networks. In order for the NSPs to provide stable network services, it is important to realize policy and QoS mechanism in the transport network. In this paper, we propose a practical method of traffic engineering cooperating IMS with MPLS (multiprotocol label switching) LSP (label switched path) selection. Our method utilizes IMS function to acquire the session profile for LSP selection. We further propose dual-phase capacity assignment, which achieves fairness and maximum resource utilization between the pairs of edge routers in our proposed architecture.

Keywords: NGN, IMS, MPLS, Traffic Engineering

1 INTRODUCTION

Many fixed and mobile NSPs (network service providers) supporting PSTN (public switched telephone networks) services are now promoting convergence on NGN (next generation network) [1] architecture, in anticipation of cost-effective synergy between legacy and Internet services. NSPs will design and construct their IP-based NGN core network to provide various services on a single network infrastructure. These services also have varying QoS requirements, for example, (a) PSTN traffic should be guaranteed, (b) some transaction or signaling/control traffic may be delay sensitive, and (c) the Internet traffic can be best effort. Nowadays, NSPs are considering more traffic accommodation in the transport stratum to provide the network resource for various services. On the other hand, customers consume more bandwidth than before. In some NSP networks, traffic required by particular customers occupies most of the bandwidth. It is therefore desirable to be able to accommodate as many customers as possible in a fair manner..

In the NGN architecture, IMS (IP multimedia subsystem) [2] is a key technology, where the CSCF (call/session control function) [3] is responsible for call (i.e. communication session) control using the SIP (session initiation protocol) [4]. NSPs can know the bandwidth demand of each session before data transmission, based on the SIP messages exchanged between UE (user equipment) and CSCFs. Such session information is transferred to the policy control server (PCRF: policy and charging rules function) in order to determine whether the session can be accepted or not. However, IMS itself does not specify the transport stratum issues, (e.g., how to realize QoS in the core transport stratum). In addition, IMS does not assume any underlying mechanism with regard to the transport stratum. On the other hand, many NSPs have introduced MPLS [5] in their transport networks to realize flexible traffic engineering, by setting up logical circuits (LSP [6]) between the pairs of edge routers reflecting various constraints and the operator's policy. In addition, NSP could collect the traffic amount per LSP, which is directly related to the pair of edge routers. This information is convenient for PCRF's call admission control to realize more precise traffic engineering. From the viewpoint of the traffic control and management facilities in NSP, it can be assumed that MPLS is often adopted in their core networks. Figure 1 abstracts the functional view of the NSP network using IMS with MPLS.

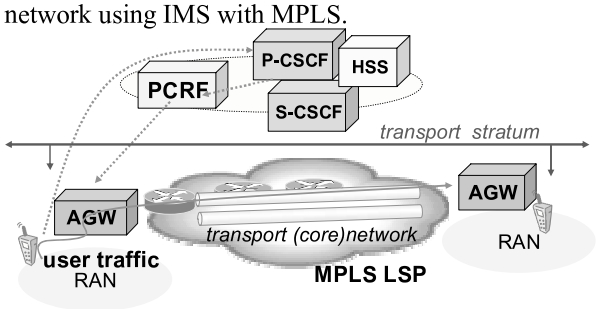


Figure1. Functional relation of NSP network using IMS with MPLS.

In this paper, we study the harmonization of IMS with MPLS-based traffic engineering for the transport stratum. Our research goal is to provide the stable communication environment to the customers, raising the traffic accommodation as well as maintaining fair resource utilization among the pairs of edge routers.

We propose an efficient MPLS LSP configuration and extension of IMS function to achieve the above goal. This paper is organized as follows. We show several issues in QoS control between IMS and MPLS in Section 2 and design the traffic engineering policy in Section 3. We explain the details of the proposed architecture in Section 4, evaluate the proposed capacity assignment method in Section 5 and show the conclusion in Section 6.

2 ISSUES ON QOS CONTROL IN NGN ARCHITECTURE

2.1 Session Control Procedure in IMS

The procedure for the call/session establishment in a mobile packet-based network is standardized in 3GPP. SIP signaling originated by UE is sent to CSCF, and the CSCF responds to the UE as to whether the session can be accepted after obtaining a decision from PCRF. The transport stratum is composed of the RANs (radio access networks), AGWs (access gateways), and transport (core) network as shown in Figure 1. The AGW is located between the RAN and the core network and enforces the QoS control for media traffic (e.g., gate opening/closing and marking the media traffic with the determined priority level). Packet marking is done by setting the bit value in the DSCP (diffserv code point) [7] or TOS (type of service) field in the IP header.

There are multiple signaling messages exchanged in establishing the session. The individual signaling messages go sequentially back and forth between UEs and CSCFs. This implies that even if the one-way delay for a signaling message takes a few milliseconds, the completion of signaling takes several times longer than sending the single message. The delay requirement for the media traffic of SIP applications generated after the signaling is less severe compared with that for signaling. Therefore, NSPs have to take into account effects to minimize the signaling traffic delay when designing and operating their networks. Now, considering the delay, the maximum domestic transmission delay (e.g., peaking at 10 milliseconds in Japan) does not give less impact on the application, but the round trip time for exchanging the signaling messages is not small. Based on these discussions, we presume that IMS-based services need, at least, the following traffic classes:

1. Class 1: both delay and loss sensitive, (e.g., signaling traffic)
2. Class 2: loss sensitive (e.g., VoIP traffic, and IPTV traffic)
3. Class 3: best-effort (e.g., Internet traffic)

2.2 MPLS Traffic Engineering

MPLS networks are composed of edge and core routers. Packets are transferred along one of the LSPs, which are established between the ingress and egress edge routers. Once a packet enters the MPLS networks (the label for LSP is assigned to the packet), core routers transfer the packet along the LSP. The label information for the LSPs among routers is exchanged by RSVP (resource reservation protocol) [8].

MPLS traffic engineering provides benefits over a plain IP network by achieving greater control of the traffic in the transport stratum. The route of an LSP can either be explicitly configured hop by hop, or dynamically routed by the CSPF (constrained shortest path first) algorithm. In a plain IP network, the shortest route to the destination is chosen by edge routers, even when it becomes more congested.

ABAF (automatic bandwidth adjustment function) has been specified and implemented [9] [10] [11] as one of the MPLS traffic engineering methods. This function not only automatically adjusts the LSP bandwidth but also dynamically reroutes the LSPs, when a certain physical link on the current LSP routes becomes short of capacity. The rerouting by the router is performed on an LSP basis; therefore, the ABAF may change the end-to-end delay of certain media traffic because of the sudden rerouting. Desirably the route is changed before the session is established. And, the operators normally want to grasp the route in the transport network. In ABAF, the route changes dynamically when the network failure does not happen. Based on these issues, we assume that it is difficult for NSP to adopt the ABAF in their MPLS networks.

From the discussion above, it is desirable that (a) the edge routers have LSPs explicitly configured, (b) have multiple LSPs for the traffic class, (c) the edge router determines the route among them for arriving traffic and disperses the traffic over their networks, and (d) the total traffic in the network is taken into account for admission. In order to meet these requirements, NSPs should have functions to recognize the demand for individual service traffic, selecting the LSP, and collecting the information on the utilization of LSPs and the physical links.

When the pair of edge routers has multiple LSPs for a specific destination, most of the procedures for traffic dispersion are conducted by the ingress edge routers, as follows:

1. Rules to distinguish media traffic and determine one of the LSPs are stored in the ingress router.
2. The incoming traffic is distinguished into one of the traffic classes with the rules
3. The packet is marked in TOS or DSCP fields of the IP header based on the destination IP address and the identified traffic class.
4. The LSP is determined with the destination IP address and the mark of the packet, and the MPLS forwarding table is looked up to find the label of the LSP.

Step 1 is realized with static rules or interworking with another entity for dynamically updating the rules. The procedure for IMS enables the dynamic updates of the rule on a session basis as described in the next subsection.

2.3 Harmonizing IMS with MPLS

MPLS is adopted in the transport stratum to explicitly control the transfer route (LSP) and to provide the fast reroute function [12]. In this paper, we consider increasing the utilization of the transport(core) network

by cooperating IMS and MPLS. IMS provides session information on UEs prior to the start of their communications. Such information is useful to determine the target LSP; however, the following items should be considered for harmonizing IMS with MPLS-based traffic engineering:

1. Deploying cooperative session control procedures between CSCFs, PCRF, and AGWs,
2. Recognizing the resource utilization of the transport stratum,
3. An admission control method to deal with multiple traffic classes,

For item 1, although the QoS/policy control architecture is being standardized in 3GPP/3GPP2 [13], QoS/policy control in the core network is largely left for the deployment. Normally we can get traffic statistics only form the physical link. But, we can get traffic statistics from the LSP using MPLS. For item 2, the LSP traffic statistics concerning resource utilization are useful. For item 3, the method should take the fairness described in section 3.2 for the accepted amount of traffic among the ingress edges into account.

2.4 Related Work

ITU-T [1] standardizes the RACF (resource and admission control function) [14] as the QoS and admission function in NGN. RACF has the same role as PCRF in 3GPP/3GPP2. However, the issue of how to adopt the RACF function to the control for the transport stratum also remains unresolved. We propose a function to control the transport stratum using MPLS.

Reference [15] numerically studied the optimal threshold for commencing traffic distribution over two LSPs and the optimal distribution rate over the two LSPs in order to maximize the admitted traffic among *n* pairs of edge routers. Their study presumed that each pair of edge routers would firstly use a single LSP between them, and secondly, they begin to use the secondary LSP when the traffic exceeds the threshold. The traffic which we want to prioritize is transferred if the threshold is exceeded. However, we deal with a sufficiently high traffic demand for each pair of edge routers and are supposedly aware of this priority. Therefore, multiple LSPs are used for traffic assignment from the beginning. We assume that the traffic demand can be recognized by tracking the traffic trend in every instance of the fixed time interval.

References [16] and [17] propose a method effectively minimizing the delay of exchanging signaling in IMS in order to ensure high communication quality. In addition, various signaling methods in IMS have been investigated in 3GPP and academic research. However, there has been little study concerning the effective treatment of signaling in terms of the traffic engineering, (e.g., the simultaneous treatment of signaling and media traffic) in IMS. Since the signaling includes various procedures during the session, low loss and latency transport network is essential to ensure communication quality perceived by users. Therefore, NSPs must take care when transferring signaling packets in the transport network.

3 DESIGN POLICY FOR TRAFFIC ENGINEERING

3.1 Traffic Class

We assume three traffic classes in our proposal, as described in Table 1. The primary class is for signaling that requires the minimum delay, while the standard class is for media traffic requiring sufficient bandwidth, as described in Section 2.1. We adopt IMS application traffic as standard class traffic. Additionally, the best effort traffic, (e.g., Internet access) without the QoS requirement, should also be taken into account.

Table 1.Traffic Class

PRIORITY LEVEL	TRAFFIC CLASS	TRAFFIC TREATMENT
High priority	Primary class	Delay restriction needed - the shortest or sufficiently small delay routes
	Standard class	Loss sensitive - traffic distribution over multiple routes to gain capacity
Low priority	Best effort	Transferred into the shortest route if there is capacity.

The additional traffic classes for more granular traffic treatment levels may be defined in certain NSPs. For example, the traffic for streaming applications requires a lower delay variation. We presume that such granular classes are treatable by the weighted round robin-based queuing discipline combined with priority queuing. We adopted at least three traffic classes in this paper to realize the traffic treatment in table 1. In this paper, a minimum of three traffic classes were used in order to validate the effectiveness of the proposed traffic engineering architecture harmonizing with IMS.

We assume that both primary and standard classes have a threshold, up to which traffic can be aggregated, while the remaining capacity can be allocated for the best effort class. This threshold (termed *acceptable capacity*) is defined for each physical link. For the bandwidth requirement, we assume that the demand for the primary class is much less than the acceptable capacity for all physical links. On the other hand, the demand for the standard class exceeds the acceptable capacity. By using multiple LSPs, and the standard class traffic can be transferred through these LSPs. For the delay requirement, the signaling traffic has a strong requirement for the minimum delay. Therefore, the

primary class traffic should take the shortest route among the LSPs between the pairs of edge routers. For the standard class, we assume the capacity requirement to be stronger than that of the delay. The standard class traffic can be distributed over the multiple LSPs.

3.2 Fairness-aware capacity assignment policy

We consider that any customer traffic should be protected at a certain minimum level. In this paper, fairness means that a certain level of capacity is guaranteed on any pair of edge routers. By providing the fairness, NSPs can raise the rate of successful connections, which leads to providing a stable communication environment to the customers.

To increase the admitted traffic for the standard class with satisfying the fairness, we focus on two traffic assignments: one is for the lowest capacity among all capacity assignments for all pairs of edge routers, and the other is for the total capacity in the transport network. In order to satisfy the above principle, we propose dual-phase capacity assignment. In this assignment, the first phase involves maximization of the lowest capacity assignment. During the second phase, the total capacity assignment is maximized from the remaining capacity on the physical links. For the first phase, an identical capacity is assigned to all the pairs of edge routers, and maximized, while for the second phase, we adopt the strategy of maximizing the total amount of the admitted traffic for the remaining capacity. The capacity of the physical link is assigned to the LSPs in the order of the smaller number of links composing each LSP. The reason for being capable of maximizing the total assigned capacity is shown as follows. When there is a certain capacity assignment A and the assignment A has an LSP that can be composed of two or more distinct and shorter LSPs, another capacity assignment, whereby the former LSP is replaced by the latter LSPs has a greater total capacity than assignment A in terms of demand and based on the pair of edge routers. Details of the method for computing the capacity assignment capturing dual-phase assignment are described in Section 5.

4 PROPOSED ARCHITECTURE

4.1 MPLS LSP Configuration

Figure 2 shows the proposed network architecture. In typical MPLS network operation, only one LSP with the shortest routes is established for every pair of (ingress and egress) edge routers. From the viewpoint of traffic accommodation, traffic engineering should be applied if detour routes are available. For example in Figure2, three LSPs along the shortest route (shortest LSPs) and an additional LSP along the detour route (detour LSP) are established between edge routers X and Y. Each LSP is assigned for the individual traffic class. One of the shortest LSPs and the detour LSP are assigned for the standard traffic class. The standard class traffic is distributed over the two LSPs.

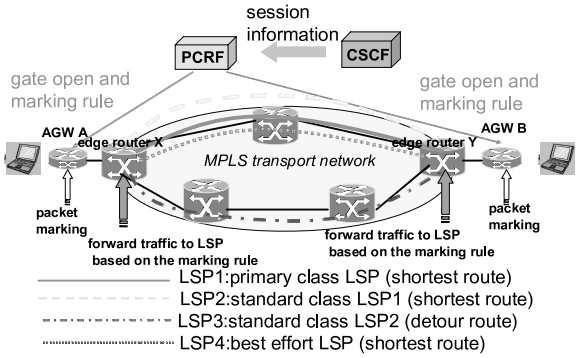


Figure 2. Proposed network architecture

The AGW filters the traffic from UEs and connects to MPLS edge routers. Each pair of MPLS edge routers has four LSPs at the minimum. LSP-1 is for the primary class traffic, while LSP-2 and LSP-3 are for standard class traffic and LSP-4 is for best effort traffic. LSP-2 should preferably be disjointed with the standard class LSP-2. We define the primary class LSP, standard class LSP-1, and the best effort class LSP as tracing the same route, namely the shortest route in most cases. In this paper, we assume the demand for the standard class to be sufficiently large and maximize the amount of standard class traffic, with fairness in mind.

The traffic utilization for each LSP unit can be collected with SNMP [18] by setting up LSP between all the pairs of edge routers. Adding the traffic of multiple LSPs between the pair of edge routers, we can know the total traffic among the edge routers. If we don't set up the LSP, we can't acquire such traffic information. The PCRF acquires that from each LSP and uses for the admission control.

In our proposed architecture, we assume that one shortest LSP and multiple detour and disjoint LSPs are statically configured. Our LSP configure does not give the additional load in the routers because the routers need not have the function like ABAF.

4.2 Bearer establishment procedure

Figure 3 illustrates the session initiation procedure in IMS standards. The AGWs set filters for QoS and policy control of the UEs' traffic. The CSCF reports the session information of the UEs to the PCRF and requests the AGWs to open the gate for the UEs' media traffic, whereupon they can start the communication through the AGWs. In our proposal, the *gate open* procedure in the AGW is expanded to perform marking

packets for the traffic engineering. Based on information from PCRF, AGW performs marking packets for the media traffic. The detailed signaling procedures, e.g. the bidirectional media-traffic treatment, between UEs, AGWs, PCRF, and CFCSs, are described in the following subsection.

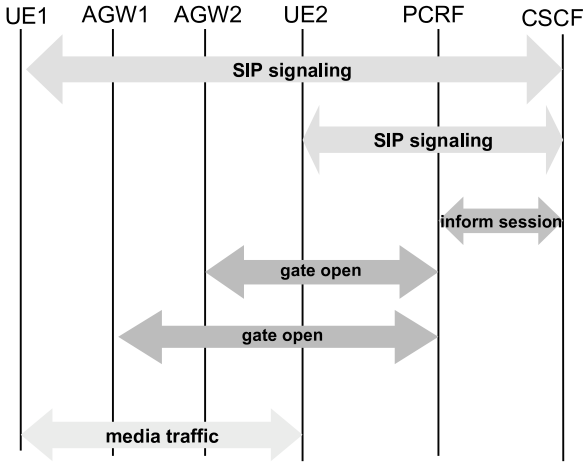


Figure 3. Session initiation procedure in IMS.

4.3 Admission Control

When media traffic session, the behavior of AGW is as follows:

1. The media traffic is transferred through the shortest LSP if the traffic is identified as the primary class.
2. If the media traffic is identified as the standard class, the traffic is distributed among multiple LSPs. The media traffic is rejected if existing media traffic in each LSP has reached the acceptable capacity for one of the links composing the LSP.
3. The media traffic identified as the best effort class is transferred through the shortest route.

To prevent packet loss for the accepted traffic of the primary and standard classes, admission control is the key for arrival session. The utilization of the physical links is regularly monitored and referenced to make a decision as to whether media traffic is accepted depending on the route to which the media traffic is transferred. The monitoring is conducted by collecting the traffic counters for all LSPs, and the utilization of all the physical links is computed. We propose that the PCRF perform the utility computation for the admission control. The PCRF can acquire the demand for arriving media traffic from the CSCFs.

We extend the standard session initiation procedure of IMS for the proposed traffic engineering, which is as follows:

1. The UE initiates the procedure with the CSCFs to establish a session.
2. The CSCF queries the PCRF to determine the LSP in the transport stratum, by which the media traffic is transferred. Here, various parameters are informed to the PCRF, (e.g., application type, IP addresses and port numbers).
3. The PCRF distinguishes the media traffic into one of the traffic classes and determines whether the media traffic can be accepted. Here, the PCRF refers to the utilization of the physical links along the LSPs assigned to the media traffic.
4. The PCRF responds to the CSCF if the LSP can be determined.
5. The CSCF requests the establishment of the bearer.
6. The PCRF sets up AGWs to open the gate for the media traffic and mark the media traffic.

The PCRF and MPLS edge routers have a common definition of the mapping between the mark (DSCP or TOS bit values) and the corresponding LSP.

The feature of harmonization between IMS and MPLS is that PCRF determines the admission and LSP selection for the media traffic with its information, which is provided by the CSCFs, and MPLS edge routers can acquire the rules used to distinguish the packets into traffic classes and marking before the media traffic arrives at the AGWs. We use the basic MPLS function, and need not expand the function of the MPLS edge router.

In the transport stratum, the traffic controlled by IMS signaling can be mixed with the Internet traffic. We assume that non-IMS-based traffic is grouped into the best effort class. So in this procedure, packets with the default mark (or no mark) are assigned to the best effort class.

4.4 LSP selection procedure

In our proposal, we presume that whenever the call of standard class traffic arrives, the PCRF determines on which standard class LSP the AGW should transfer the media traffic and whether it is accepted or not. The PCRF calculates the capacity assignment for media traffic acceptance beforehand. The utilization of the physical links and LSPs, which the PCRF recognizes, is updated at specific time intervals (e.g., every 1 or 3 minute/s), since the traffic counter values for LSPs are collected at this interval. The capacity assignment for the media traffic acceptance for all the LSPs is also updated at this interval. The LSP selection for the standard traffic class is also conducted in proportion to the assigned capacity. On the other hand, the PCRF responds *cancel* to the CSCFs if any of the media traffic (e.g. the VoIP service using bidirectional traffic) is reject.

5 CAPACITY ASSIGNMENT

5.1 Modeling of dual-phase capacity assignment

We applied the LP (linear programming) approach to achieve the goal of our research. Here, we presumed that the demand was a real number to simplify the computation of the LP. We considered the model for dual-phase capacity assignment for the standard class traffic: the first assignment maximizes the minimum capacity among all the pairs of edge routers, with fairness in mind; and the second assignment maximizes the total capacity for the remaining capacity in the transport stratum.

Maximizing the minimum capacity is computed by solving the LP, which maximizes the identical capacity assigned to all the pairs of edge routers. We define the following objective function for the first phase assignment:

$$C = d_k = \sum_i d_{k,i}$$

where d_k, $d_{k,i}$ denote the assigned total capacity of the edge router pair *k*, and the capacity of LSP *i* for edge router pair *k*. Here, we define the following constraint conditions for the above objective function:

$$\sum_k \sum_i x_{e,k,i} - u_e \le 0 \text{ for } e \in E,$$

where $x_{e,k,i}$ and u_e denote the assigned capacity of LSP *i* for the edge router pair *k* in the physical link *e*, and the available capacity for the standard class traffic in the physical link *e*. Although identical capacities are assigned to each pair of edge routers, this capacity can be distributed via multiple LSPs of each pair of edger routers.

Similarly, the second phase assignment is also computed by solving the LP, which maximizes the total capacity $C = \sum_k \sum_i f_{k,i}$ under the constraint conditions $\sum_k \sum_i y_{e,k,i} + \sum_k \sum_i x_{e,k,i} - u_e \le 0$ for $e \in E$, where $f_{k,i}$ and $y_{e,k,i}$ denote the additional assigned capacity of LSP *i* for the edge router pair *k*, and that in the physical link *e*.

5.2 Evaluation Method

To evaluate the effect of the proposed capacity assignment in PCRF, we performed the simulation and compared the average and worst admitted in all pairs of edge routers between four traffic assignments by varying the number of routers in the MPLS network (edge and transit routers). The first is the dual-phase capacity assignment with single LSP for the standard class traffic in each pair of edge routers. This is termed "1-*path max-min*". The second is the dual-phase capacity assignment with two LSPs in each pair of edge routers (termed "*2-path max-min*"). The third is the dual-phase capacity assignment with three LSPs in each pair of edge routers (termed "*3-path max-min*"). For the fourth, only the second phase of the dual-phase capacity assignment is applied without any fairness consideration. It has the two LSPs for each pair of edge routers, and is termed "*2-path shortest-first* ".

To emulate the NSP topology, we used BRITE (Boston university Representative Internet Topology) [19]. BRITE is a useful tool for emulating network topology in an AS (autonomous system). BRITE provides the BA (Barabasi-Albert) [20] model, which is often used to emulate the topology. We specified the number of routers and the degree, (the number of physical links per individual router) and generated network topologies for the simulation.

We specified that all links in the generated network topology had identical link capacities. In addition, the shortest LSP and detour LSPs (when multiple LSPs were used) were computed in all the pairs of edge routers. We need to be careful of the crossover of some detour LSP, when setting up the detour LSP in the network topology.

In generating the network topology, we defined 80% of the routers as edge routers, and 20% as core (transit) routers. This reflects the NSP's MPLS networks, where there are a larger number of many edge routers which are the head and tail ends of the LSPs, and a smaller number of core routers, which transit traffic for edge routers by switching the LSP. In the simulation, the edge routers had a degree of at least "2" since the edge router normally has two physical interfaces to connect the core (upper) network in the NSP. The core routers had a degree of over three.

The four capacity assignments were compared in the same network topology provided by BRITE, while BRITE also varies the topology as it generates. We tried ten simulations per the number of routers, and the results were averaged. To solve the modeled formula of LP, we use GLPK (GNU Linear Programming Kit).

5.3 Evaluation Result

The results are shown from Figures 4 to 7, with Figures 4 and 5 showing the average and worst amounts of traffic admitted at all the edge routers, respectively. The X-axis is the number of edge routers, while the Y-axis is the average or worst volume to link capacity, and indicates the admitted traffic to link capacity in edge routers. The value is normalized by the link capacity.

The average amount of the admitted traffic to link capacity (Figure 4) shows that the 2-path shortest-first assignment achieves the largest traffic accommodation. But, compared with 2-path max-min assignment and 2-path shortest-first assignment, although 2-path max-min assignment considers the fairness, the averages amount of admitted traffic of 2-path max-min assignment is more 80% than one of 2-path shortest-first assignment. Comparisons among the 1-, 2- and 3-path max-min assignments reveal that the traffic accommodations become increasingly similar, as the number of edge routers increases.

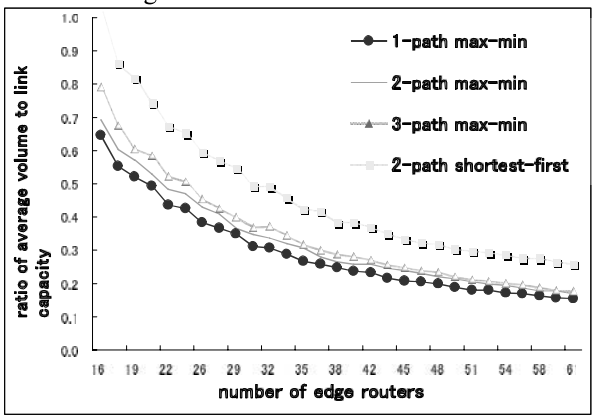


Figure 4. Average amount of traffic admitted in edge routers.

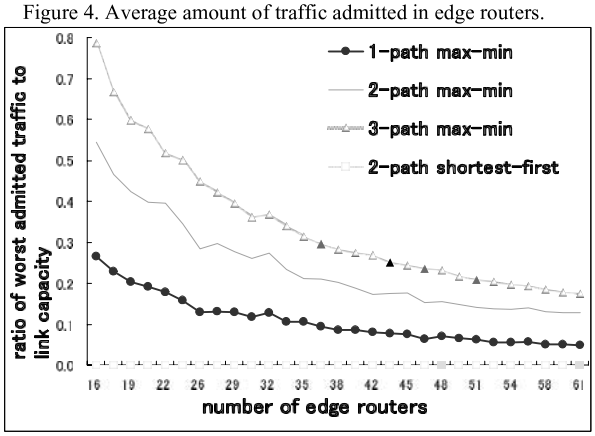


Figure 5. Worst amount of traffic admitted in each edge routers.

In the case of the worst (lowest) accommodation (Figure 5), the 3-path max-min assignment achieves the largest traffic accommodation. The 2-path max-min assignment admits traffic almost twice as large as the 1-path max-min. For 3-path maxi-min assignment, the admitted traffic is 1.2-1.3 times larger than the case of 2-path max-min assignment.

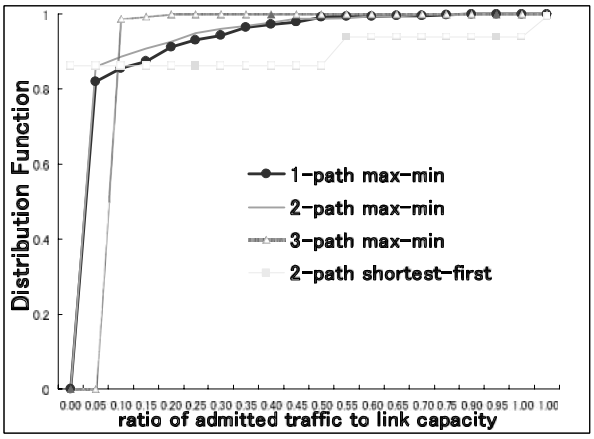


Figure 6. Cumulative distribution function in 20 edge routers

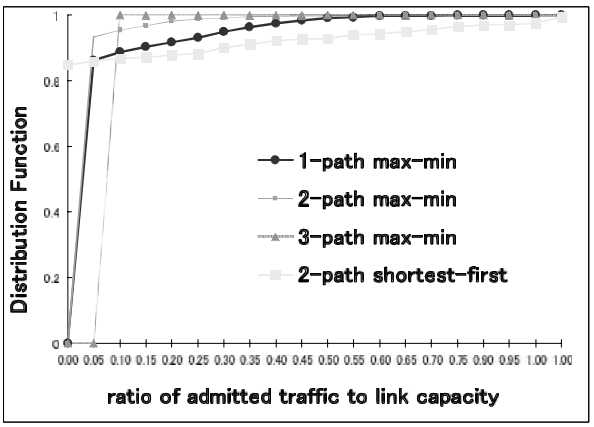


Figure 7. Cumulative distribution function in 40 edge routers.

Figure 6 and 7 show the cumulative distribution functions for the ratio of admitted traffic to link capacity in the case of 20 and 40 edge routers, respectively. These figures show that relatively large amount of edge routers fail to admit any of the demand for the 2-path shortest-first assignment. The results of the 2-path shortest-first assignment indicate that some of the edge routers can not admit any traffic demand. For the other assignments, the fairness of the traffic accommodation is achieved.

In terms of the fairness accommodation, it is severe to take the 2-path shortest-first assignment in the NSP's operation and service. These results indicate that the 2-path shortest-first assignment is impractical for NSP's operation whereas 2- and 3-path max-min assignments are effective in terms of fairness.

Focusing on the number of LSPs in each pair of edge routers, the progress of 3-path max-min assignment from 2-path max-min assignment is less than that of 2-path max-min assignment from 1-path max-min assignment, while its progress becomes more remarkable as the number of edge routers increases. This is probably because the numbers of disjoint LSPs for 3-path case are similar to those for the 2-path case. Moreover, this limitation may result from the number

of degrees (two for edge routers and three for transit routers) in the simulation network topology. Generally, it is more difficult to set up multiple disjoint paths as the number of LSPs in each pair of edge routers increases. We presume that the progress of 4-path max-min assignment or cases using more paths is probably much less than that of 3-path max-min assignment from 2-path max-min assignment if the simulation had the same topology.

The difference of each value in Figure5 among the value of the 1-,2-,3-path max-min assignment is more than one in Figure 4. We can say that the worst value of capacity assignment is remarkably praised by setting more multiple LSPs from the results. Setting more multiple LSPs in the pairs of edge routers is more effective to achieve the fairness accommodation than to utilize the network resource.

The number of variables in LP to solve the capacity assignment increases as the number of routers increases in the simulated topology. However, even if the number of edge routers was over 60, in our simulation, the computation time required to solve the modeled LP took about less than one second. We used an off-the-shelf PC with Intel Core™ 2 CPU of 2.00 GHz and 1.99GB as Memory for the simulation. Therefore, the computational load of our proposed capacity assignment is sufficiently low.

5.4 Discussion

We set "2" as the value of the least degree in BRITE, and simulated the network topology of NSP. We need to examine "1" case or a bigger value case. If we set "1" as it, we assume the effect among the 1-,2-,3-path max-min may be little. If we set a bigger value as it, we assume that the progress of 4-path max-min or cases using more LSPs may be more and we can achieve the more utility of network resource. However, taking account into the realistic network topology of NSPs, edge routers usually has two physical interfaces for redundant access to core routers. So, it is enough to consider the least degree "2" case in the simulated topology.

Additionally, increasing the number of LSPs for each pair of edge routers brings, to the network operation, complexity in terms of the LSP maintenance, e.g. recovering preparation for the failure and utility monitoring. The number of LSPs set up in the pairs of edge routers includes a trade-off between the operational cost and the resource utilization. One future issue involves performing validation, and increasing the number of degrees in the network topology.

6 CONCLUSION

In this paper, we proposed the architecture for harmonizing IMS with MPLS-based traffic engineering, reflecting the objective of providing the stable service to the customers in NGN infrastructure. We showed the extended PCRF function to utilize IMS session information for resource management of the MPLS core network. IMS provides the session information, of the media traffic and MPLS utilizes that information to perform traffic engineering. We presented the benefits of harmonizing IMS with MPLS and proposed the required functions, architecture, and procedure. To implement our proposal, the MPLS routers need not have additional functions. To achieve that objective in our work, more various methods can be generally thought. But, in this paper, first of all, we propose that for the signaling traffic is prioritized in primary class LSP and the media traffic is transferred to multiple standard class LSPs. With regard to traffic engineering in the proposed architecture for the standard class traffic, we proposed the dual-phase capacity assignment to maximize the lowest value in the capacity assignment, and maximize the remaining bandwidth for the standard class traffic. In the evaluation of the capacity assignment, we compared the effect of our proposed capacity assignments with the 2-path shortest-first, which is equivalent to the second phase of the proposed capacity assignment. As the result, we showed that the worst amount of traffic admitted in the edge routers increased, corresponding to the number of LSPs per router pair. However, the average amount of traffic admitted in the edge routers when adopting one, two and three LSPs was almost identical, regardless of the number of edge routers.

REFERENCES

[1] ITU-T NGN release1 Y.2001 "General overview of NGN"

[2] ETSI "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN)", ETSI E2 282 001 V1.1.1 NGN Functional Architecture Release 1.

[3] ETSI, "Telecommunications and Internet Converged Services and Protocols for Advanced Networking (TISPAN); IP Multimedia Call Control Protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP) Stage 3", ETSI ES 283 003, V1.1.1, April 2006.

[4] M. Handley, H. Schulzrinne, E. Schooler and J. Rosenberg, "SIP: Session Initiation Protocol", IETF RFC 3261, June 2002

[5] E. Rosen, A. Viswanathan, R. Callon, "Multiprotocol Label Switching Architecture", IETF RFC3031, January 2001

[6] D. Awduche, J. Malcolm and J. Agogbua, "Requirements for Traffic Engineering Over MPLS", IETF RFC2702, September 1999

[7] K.Nicols, S.Blake, F.Baker, " Definition of the Differentiated Services Field (DS Field)in the IPv4 and IPv6 Headers", IETF RFC2474, December 1998

[8] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan and

G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", IETF RFC 3209, December 2001.

[9] Aumatic Bandwidth Adjustment for TE tunnels, ”QoS for IP/MPLS networks” Cisco Press ISBN 1587052334

[10] H. Ishi, K. Nagami, ”Proposal and evaluation for Automatic bandwidth setting in MPLS LSP” IEICE B Vol. J89-B No. 10 pp.1894-1901.

[11] F.Le Faucheur, ”Aggregation of RSVP(Resource ReSerVation Protocol) Reservations over MPLS TE/DS-TE Tunnels” IETF RFC4804, Feburary 2007.

[12] P. Pan, G. Swallow, A. Atlas, ”Fast Reroute Extensions to RSVP-TE for LSP Tunnels”, IETF RFC4090, May 2005

[13] 3GPP2 X.S0013, “All-IP Core Network Multimedia Domain”.

[14] ITU-T NGN release1 Y.2111

[15] H. Tamura, K. Kawahara, Y. Oie, ”Analysis of two-phase Path Management Scheme for MPLS Traffic Engineering” Proc. Of SPIE Performance Quality of Service, and Control of Next Generation Communication Networks2 pp. 194-203, Oct. 2004.

[16] S. Zaghloul, A. Jukan, W. Alanqar, "Extending QoS from Radio Access to an All-IP Core in 3G Networks: An Operator's Perspective", IEEE Communications Magazine.

[17] F. Wegscheider, "Minimizing unnecessary notification traffic in the IMS presence system", 1st International Symposium on Wireless Pervasive Computing, 2006.

[18] C. Srinvasan, Bloomberg L.P.,"Multiprotocol Label Switching (MPLS) Traffic Engineering (TE) Management Information Base (MIB)", IETF RFC 3812, June 2004.

[19] A. Medina and I. Matta, “BRITE:A Flexible Generator of Internet Topologies”, Tech. Rep. BU-CS-TR-2000-005, Boston University, Boston MA, 2000.

[20] R. Albert and A. Barabasi, “Statistical mechanics of complex networks”, Review of Modern Physics, 2002.

Determining the Relay Node Encode Packet in Multipath Routing Environment

Tomonori Kagi [*] and Osamu Takahashi [**]

[*] Graduate School of Informatics, Future University-Hakodate, Japan
[**] Faculty of Informatics, Future University-Hakodate, Japan
{g2107006, osamu}@fun.ac.jp

Abstract - Wireless networks, such as mobile ad hoc ones have low reliability for reasons such as phasing, noise, and packet collisions. FEC-based methods in ad hoc networks have been improved because of their increased reliability when used with multipath routing. However, the number of transmission packets from the source node has been increased. Therefore, we propose using an efficient and reliable packet transmission method: using multipath routing constructs from multiple node disjoint routes and applying network coding, which allows packet encoding at a relay node. Because the encoding packet is generated by a relay node, the source node does not need to encode the packets, and it sends only unencoded packets to each route. Thus, the number of packets transmitted by the source node does not increased. In addition, we also evaluated which node was most suitable to encode a packet, the location of the path that should be used to encode it and the delivery ratio by the number of packets used for encoding.

Keywords: Wireless network, MANET, Network Coding, Multipath Routing

1 INTRODUCTION

Recently, progress in wireless communication technology has meant that wireless modules have been mounted on various devices. These ad hoc networks are instantly deployable wireless networks, which rely on radio waves instead of base stations or communication infrastructure support. Because radio waves have a short propagation range, the route becomes "multihop" when a communication peer is not within range. In general, the reliability is low in ad hoc networks because of network topology, unstable radio environment, and packet collisions. By "reliability" we mean the probability that data generated at a source node in the network can be routed to the intended destination.

Packet-level forward error control (FEC) and automatic repeat request (ARQ) are two methods widely used to recover the lost packets in networks with unreliable links.

Automatic repeat request is an error recovery method that uses acknowledgment packets (ACKs) and a timer to transmit data reliably. The acknowledgment packet is a message sent by the receiver to the sender to indicate that it has correctly received a packet. If the sender does not receive an acknowledgment before a specified period of time (timeout), the sender usually retransmits the packet until it receives an acknowledgment or exceeds a predefined number of retransmissions. However, the ARQ method is not considered applicable in networks that have low reliability and that are highly mobile, such as ad hoc ones. This is because the transmission delay increases as a result of retransmissions by the sender for missing ACKs. In addition, because of its use of unidirectional links, ARQ is unfit for wireless networks [5].

Forward Error Control is an error correction method that is used in data transmission in which the sender generates an error correction code, adds it to the original packet, and then sends both the error correction code and the original packet. Using this method allows the receiver to detect and correct errors without the need to ask the sender for retransmission of the packet.

The use of FEC-based methods in ad hoc networks has been studied [6] [7] [8] and found to improve the reliability when used with multipath routing. However, the number of transmission packets of the source node is increased. For example, as shown in Figure 1, the source node S generates a code from Data 1 and Data 2 by encoding them. The source node then sends the code. In this case, the number of packets transmitted by the source node is three (Data 1, Data 2, and Code). Thus, the transmission frequency at the source node is increased.

We proposed using a method that involves Network Coding [2] that allows packet encoding at a relay node to decrease the number of packet transmitted by a source node and the number of packets that flow into the network accompanying it. In addition, we carried out a computing simulation to evaluate it [1].

The results from the simulation show that our proposed method transmits data more efficiently and more reliably than the current method does. However, when multiple paths are constructed from multiple relay nodes, how to decide on which node should be encoded has yet not been discussed. Thus, we theoretically evaluated which relay nodes should encoded and also on which paths packet from the delivery ratio and packet overhead, and the delay. In addition, to validate our theoretical evaluation, we conducted simulations.

Our method is discussed in Section 2. A prototype implementation of our proposal is described in Section 3. We evaluated our proposal by comparing related

protocols in Section 4, and we summarize our work in Section 5.

2 PROPOSED METHOD

In this section, we describe our proposed method using Network Coding with Multipath Routing.

2.1 Basic Operation Model

The construction of the multiple route method is Split Multipath Routing [4], which is also known as extended Dynamic Source Routing [3].

Multiple paths are constructed, and then the source node sends data packets to neighbor nodes on all routes simultaneously. For example, as shown in Figure 1, when two paths are constructed a data packet is forwarded on one path, and a certain relay node on another path encodes a packet and forwards it.

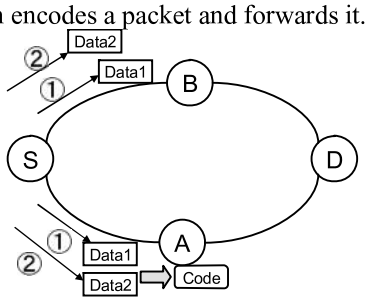


Figure 1 Proposed method

Therefore, the destination node is able to receive encoding packet even if there is no source node that encodes the packet and then transmits the encoding packet.

3 CHOICE OF NODE

Our proposal method is modeled by using a mathematical expression.

In addition, we evaluated whether a data packet should be forwarded and the position of the path where the encoding packet should be generated and forwarded. These conditions were on the basis of a theoretical formula.

Similarly, we evaluated which relay node should encode a packet, when the path is constructed by two or more relay nodes.

In order to estimate the efficiency of our proposal, we evaluated the packet delivery ratio, the packet overhead, and the transmission delay. We define these parameters as follows.

- Packet delivery ratio

The packet delivery ratio is defined as the number of correctly received data packets at the destination node divided by the number of original data packets sent by the source node.

- Packet overhead

The packet overhead is defined as the number of all node transmission packets, including data packets and encoded packets.

- Transmission delay

The transmission delay is defined as the period from which the source node generates a packet until the time when the destination node receives it. For encoding models, the transmission delay is defined as the period from when the source node generates a packet until the time the destination node decodes the encoded packets and retrieves the original ones.

3.1 Evaluation Model

The evaluation model is shown in Figure 2

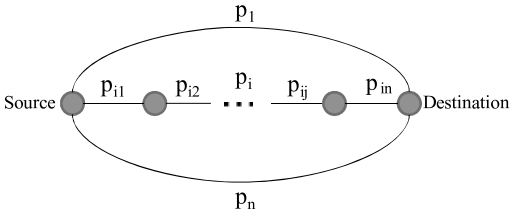


Figure 2 Evaluation model

Packet Loss Rate between nodes in the *i*th path is given by:

$$(p_{i1} \quad p_{i1} \quad \cdots \quad p_{ij} \quad \cdots \quad p_{in}) \qquad (1)$$

Packet Loss Rate between the source node and the destination node in each path is given by:

$$(p_1 \quad p_2 \quad \cdots \quad p_i \quad \cdots \quad p_m) \qquad (2)$$

The number of hops between the source node and the destination node in each path is given by:

$$(H_1 \quad H_2 \quad \cdots \quad H_i \quad \cdots \quad H_l) \qquad (3)$$

The packet loss rate in each path, which is computed from the packet loss rate between nodes, is given by:

$$p_i = 1 - \prod_{j=1}^{Hi} (1 - p_{ij}) \qquad (4)$$

3.2 Coding Scheme

When we constructed the redundancy packets, we assumed the that the encoding we used would have parameters (N, K, and t), where (i) N is the number of transmitted packets in a group; (ii) K is the number of the data packets in this group; and (iii) t is the erasure recovery capability, i.e., the maximum number of lost packets within the group that can be reconstructed on the basis of the received packets.

3.3 Precondition

In order to simplify the evaluation, the following items are defined as a precondition.

- Number of routes: 2

- Transmission path for data packets: path1
- Transmission path for encoding packets: path2
- Encoding parameters: (*N*, *K*, 1)

3.4 Packet Delivery Ratio

The relay node, which encodes is determined for the better packet delivery ratio. The following parameters are defined.

One of the following two requirements needs to be filled in order to send a packet to destination node correctly.

(Requirement 1) Destination is received data packet.

(Requirement 2) Destination node is received packet required decoding.

The probability of meeting requirement 1 is given by:

$$P_1 = 1 - p_1 \quad (5)$$

We then calculated the probability of meeting requirement 2.

The encoding packet is generated and delivered to a destination node.

The probability that the source node can transmit one packet to the relay node that is an encoding packet is given by:

$$P_{e1} = \prod_{i=1}^{H_e} (1 - p_i) \quad (6)$$

The variable *He* is the number of hop to a relay node, which encodes the packet.

The probability that the destination node receives two, three, ..., K packets required for coding is approximated by:

$$P_{e2} = \sum_{n=0}^{X} \prod_{j=1}^{H_e} (1 - (1 - p_{2j}))^n (1 - p_{2j}) \approx 1 \quad (7)$$

The probability that the source node can transmit a packet required for encoding to the relay node, which is encoding packet is given by:

$$P_e = P_{e1} P_{e2} = \prod_{i=1}^{H_e} (1 - p_i) \quad (8)$$

The relay node encodes packets if the packet required for encoding is received.

The probability that the relay node can encode and forward the packet to the destination node is given by:

$$P_{ed} = \prod_{j=H_e+1}^{H_2} (1 - p_{2j}) \quad (9)$$

From (8) and (9), the probability that the encoding packet is generated by a source node is forwarded to the destination node is given by:

$$P_{sd} = P_e P_{ed} = 1 - p_2 \quad (10)$$

The probability data packet is decoded from an encoding packet and another data packet required for decoding without receiving a data packet is given by:

$$P_2 = p_1 (1 - p_1)^{K-1} P_{sd} = p_1 (1 - p_1)^{K-1} (1 - p_2) \quad (11)$$

On the basis of (5) and (11), the packet delivery ratio is obtained by:

$$P = 1 - p_1 + p_1 (1 - p_1)^{K-1} (1 - p_2) \quad (12)$$

(1) Determination of relay node encode

We evaluated which relay node should be encoded.

Formula (12) does not contain the variable *He*. Thus, we were able to set up a packet delivery ratio regardless of the number of hops to a relay node that encodes the packet.

The packet delivery ratio is the same regardless of node encodes packet.

(2) Determination of the Path

Next, we evaluated whether a data packet should be transmitted on the path and in which conditions.

The formula which is transformed from (12) is given by:

$$P = 1 - p_1 \{1 - (1 - p_1)^{K-1} (1 - p_2)\} \quad (13)$$

For both two possible relative values the path $p_1 < p_2$ or $p_1 > p_2$, we evaluated, the packet delivery ratio increases.

If the path in which p_1 become smaller is determined, $(1 - p_1)^{K-1} (1 - p_2)$ increases. Therefore, P increases as p_1 decreases. Thus, we determined the path such that $p_1 < p_2$. If the packet loss rate is defined as equal among all the nodes, the path which forwards data packet is determined such that $H_1 < H_2$. Thus, we determined which data packet is forwarded on the path with fewer hops.

3.5 Packet Overhead

Next, we evaluated the packet overhead.

Packet Overhead O is set by setting X as all data packet, which should be sent is given by:

$$O = \frac{X}{K} \left\{ K \left(\sum_{j=1}^{H_1} (1 - p_{1j})^{j-1} + \sum_{j=2}^{H_e} (1 - p_{2j})^{j-1} \right) + \sum_{j=H_e+1}^{H_2} (1 - p_{2j})^{j-1} \right\} \quad (14)$$

(1) Determination of relay node encode

We decided on the basis of our evaluation, which relay node should be encoded.

We see that the packet overhead O decreased as the amount of *He* decreased as a result of (14). Thus, the relay node adjacent to the source node should encode the packet.

(2) Determination of the path

We evaluated whether a data packet should be transmitted on the path and in which conditions from number of hops.

Packet overhead increases because the number of packets that are forwarded decreases as the packet loss rate increases.

However, packet delivery ratio decreases as the number of packets that is forwarded decreases. Therefore, we did not find the packet delivery ratio because it is factor.

We see that O decreases when the path is determined such that $H_1 < H_2$. Thus, the data packet should be forwarded on the path in which number of hops is fewer.

3.6 Delay

To simplify the evaluation, the following items are defined as assumptions.

We evaluated the delay on the basis of the following parameters.

t_s : transmission interval[s]

t_n : wireless delay[s]
(fixed among all nodes)

The average delay, $T_{s,}$ which forwards a data packet to a destination node is given by:

$$T_s = H_1 t_n (1 - p_1) \quad (15)$$

The average delay, $T_{e,}$ the period from when a data packet is decoded from an encoding packet to another data packet is given by:

$$T_e = p_1(1-p_2)(1-p_1)^{K-1}\prod_{j=1}^{H_e}(1-p_{2j})(K-1) \times\sum_{i=1}^{X}((\max(H_1,H_2)t_n + t_s i)(1-\prod_{j=1}^{H_e}(1-p_{2j}))^{i-1}) \quad (16)$$

On the basis of (15) and (16), average delay is obtained by:

$$T = \frac{T_s + T_e}{P} = \frac{p_1(1-p_2)(1-p_1)^{K-2}(K-1)\left(\max(H_1,H_2)t_n + t_s\prod_{j=1}^{H_e}(1-p_{2j})^{-1}\right) + H_1 t_n}{p_1(1-p_2)(1-p_1)^{K-2}+1} \quad (17)$$

(1) Determination of relay node encoding

We evaluate which relay node should code.

We see that the delay T decreases because $\prod_{j=1}^{H_e}(1-p_{2j})^{-1}$ decreases as He decreases since (17).

Thus, the relay node adjacent to the source node should encode the packet.

(2) Determination of Path

Next, we evaluated whether the data packet should be forwarded on the path and in which conditions.

(16) with H_e=1 is given by:

$$T = \frac{p_1(1-p_2)(1-p_1)^{K-2}(K-1)\left(\max(H_1,H_2)t_n + t_s\right) + H_1 t_n}{p_1(1-p_2)(1-p_1)^{K-2}+1} \quad (18)$$

We evaluated the delay only when the number of hops was changed.

We assumed that p1 and p2 do not change even if the number of hops changes.

For either values of H, i.e., when $H_1 > H_2$ or when $H_1 < H_2$, we found that the delay decreases. In either case, T remains unchanged because max(H_1, H_2) does not change. Therefore, the delay T decreases as H1 decreases. Thus, the data packet should be forwarded on the path that has fewer hops .

The path on which the data packet is forwarded does not determine the change of the packet loss rate because of the dependence on the number of hops and on each delay time.

4 EXPERIMENT

We verified the validity of the result of the theoretical evaluation by carrying out a computor simulation. We used ns2, a discrete event simulator, [9]. The simulation topology is shown in Figure 3.

The simulation environment is shown in Table 1.

We only considered packet loss from data packets when we evaluated the data transmission rate

Packet Loss Rate is defined as being an equal value among all the nodes.

Table 1 Simulation parameters

Field [m]	1000 × 1000
Number of Nodes	7
Radio range [m]	250
Speed [km/h]	0
Simulation time [sec]	500
Data size [bytes]	512
Transport Protocol	UDP
Time between generating packet [s]	0.25
Packet Loss Rate [%]	0 ~ 50
Encoding parameter	(3,2,1)

4.1 Simulation results

In three evaluation models shown in Table 2, the theoretical result obtained from the formula and the simulation result are shown in Figure 6 through Figure 4.

Table 2 Evaluation model

	Path data packet is forwarded	Path encoding packet is forwarded	He
Model 1	Path 1	Path 2	1
Model 2	Path 1	Path 2	3
Model 3	Path 2	Path 1	1

・Packet Delivery Ratios

The results obtained from Formula (12) (Theory) and the simulation result (Simulation) are shown in Figure 4.

To determine the relay node that encodes the packet, the packet delivery ratio remains unchanged regardless of the number of hops to a relay node which encodes the packet as well as the result proven by the formula.

To determine the path on which the data packet is forwarded, Model 1 has a higher packet delivery ratio than that of Model 3. Thus, this proves that the method to forward data packet should be forwarded on the path in which the packet loss rate is low (the path with few hops) has higher packet delivery ratio.

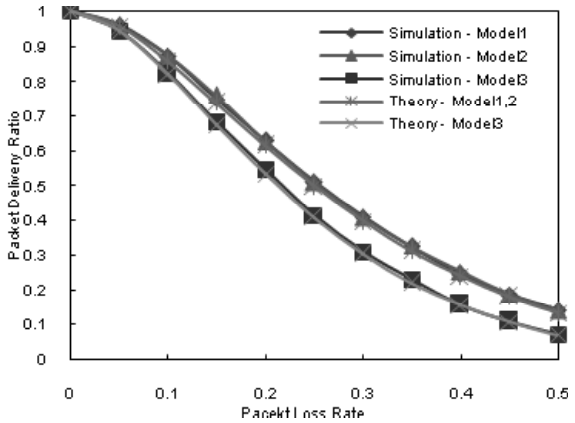


Figure 4 Packet Delivery Ratio

・Packet Overhead

Figure 5 shows the packet overhead.

The result obtained by (14) is the same as that obtained by using a simulation.

Model 1 has the smallest, and Model 3 has the largest packet overhead. Thus, we proved that the way to forward a data packet is to use the path in which the total number of hops is low and the number of hops to the relay node that encodes the packet is low has lower packet overhead.

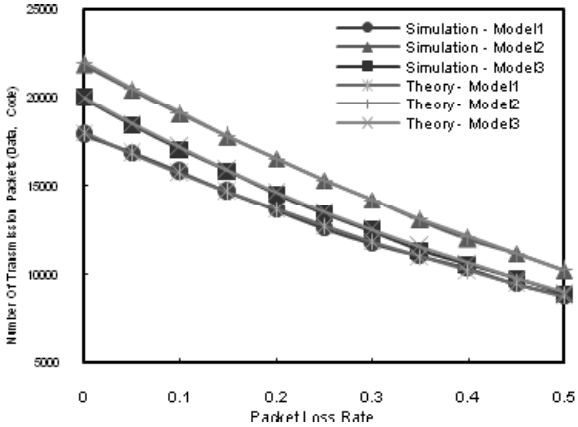


Figure 5 Packet Overhead

・Delay

The delay is shown in Fig. 6: the graph shown was obtained by using Formula (17) with $t_s = 0.25$ and $t_n = 0.01$ and the simulation result.

The graph obtained by using the formula (Theory) is not equivalent to the simulation result (Simulation) because we did not include the time needed to construct a path nor the validity of the set-up wireless communication delay between each node.

However, the result for the magnitude relation is same as that obtained by the simulation, i.e., by determining the path and the relay node encode.

Model 2 has a higher delay in comparison with that obtained by using Model 1. Thus, the delay increases with the number of hops to the relay node that encodes the packet. As a result, the relay node nearest the source node encodes the packet.

In addition, Model 3 has a higher delay than Model 1. Thus, this proved that the delay decreases when a data packet is forwarded on a path with fewer hops.

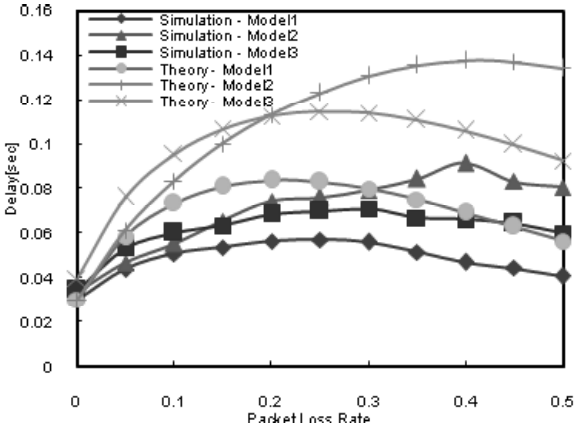


Figure 6 Delay

4.2 Summary

The data packet that should be forwarded on a particular path and in which conditions, and which relay node should encode a packet is shown in the results of the evaluation (Table 3).

Our method has a higher packet delivery ratio, a lower packet overhead.

The number of hop to a relay node that encodes packet has no relation to the packet delivery ratio. Furthermore, the relay node adjacent to the source node should encode the packet.

Table 3 Result

	Packet Delivery Ratio	Packet Overhead	Delay
Path data packet is forwarded (Packet Loss Rate)	Low	-	-
Path encoding packet is forwarded (Hop)	Few*	Few	Few
Number of hops to a relay node which encodes packet	independent	1	1

*When Packet Loss Rate between all nodes is equal.

5 CONCLUSION

We evaluated our proposed method by using network coding with a multipath routing environment.

We evaluated only two paths: the first path in which data packet is forwarded, and the second one in which the encoding packet is forwarded and limited by encoding parameters with (N, K, 1).

To determine the path on which the data packet is forwarded, we proved that the data packet should be forwarded on a path with few hops when the packet loss rate is equal among all nodes.

To determine which relay node should encodes the packet, we theoretically proved that the relay node adjacent to the source node encodes packet is better.

In addition, we proved the validity of the theoretical evaluation with a simulation.

6 FUTURE WORK

We evaluated a limited number of paths. In addition, all data packets and encoding ones are distributed and forwarded in each path.

However, to ensure that our work takes load balancing into consideration we should evaluate how a packet should be scheduled when more than three paths are constructed.

In addition, we proved that the packet delivery ratio is better when a data packet is forwarded on the path in which the packet loss rate is lower. However, the path is not determined if this rate is not measured. Thus, we need to investigate how to measure the Packet Loss Rate on each path.

REFERENCES

[1] T. Kagi, O. Takahashi, " Efficient Reliable Data Transmission using Network Coding in MANET Multipath Routing Environment", IPSJ, MBL44, pp.209-216(2008)

[2] R. Ahlswede, N. Cai, S. R. Li, and R. W. Yeung. "Network information flow", *IEEE Transactions on Information Theory*, (2001)

[3] D. Jhonson, Y. Hu, and D. Maltz, "The Dynamic Source Routing Protocol (DSR) for Mobile Ad Hoc Networks for IPv4", IETF-Request-for-Comments, rfc4728.txt. (2007).

[4] S. J. Lee, and M. Gerla "Split Multipath Routing with Maximally Disjoint Paths in Ad hoc Networks", ICC, pp. 3201-3205(2001)

[5] M. Kosugi, H. Higaki, "Data Message Transmission in MANET with Uni-Directional Links", IPSJ, MBL41, pp.103-106(2007)

[6] W. Lou, W. Liu, and Y. Zhang, "Performance Optimization using Multipath Routing in Mobile Ad Hoc and Wireless Sensor Networks", *Combinatorial Optimization in Communication Networks*, edited by M. Cheng, Y. Li, and D-Z. Du, Kluwer, (2006)

[7] R. Ma and J. Ilow, "Reliable Multipath Routing with Fixed Delays in MANET Using Regenerating Nodes," In IEEE LCN'03, (2003).

[8] R. Ma and J, Ilow, "Regenerating Nodes for Real-Time Transmissions in Multi-Hop Wireless Networks," In IEEE LCN'04, (2004).

[9] ns-2. http://www.isi.edu/nsnam/.

A Study on Integrated Protocol for Communications and Positioning

Tomoya Takenaka[†], Hiroshi Mineno[‡], and Tadanori Mizuno[†]

[†]Graduate School of Science and Technology, Shizuoka University, Japan
[‡]Faculty of Informatics, Shizuoka University, Japan
{tomoya, mineno, mizuno}@mizulab.net

Abstract - The purpose of this paper is to introduce an integrated protocol for communications and positioning. The motivation for designing this integrated protocol comes from the recent research results for the location information. The objective of the integrated protocol is to enable simultaneous data exchange and location discovery. We describe the protocol stack of the integrated protocol and look at the facilities required at each layer. For the MAC layer, resource control for positioning is introduced. For the NWK layer, simultaneous localization and routing is discussed. Target tracking, an application of the integrated protocol is also investigated.

Keywords: positioning, data communications, wireless multi-hop networks, ad-hoc networking

1 Introduction

Emerging wireless networking capabilities and micro-electronics technologies enable the provision of the various types of networks, such as ad-hoc and sensor networks. Zigbee [1] is one of the emerging standardized sensor network products. Once sensor nodes are deployed, they can gather the sensing information for a observer automatically. In such sensor networks, sensing data is expected to be bundled with location information to locate the event. To estimate node positions, positioning techniques have been discussed in areas such as cellular communications [3] and wireless multi-hop networks [4].

The relationship between data communication protocols and positioning protocols is presented in Fig. 1. Location information is not only used for sensor networks, but also for improving networking performances. Location information enables the reduction of redundant packets [12], [13]. In location-aided routing (LAR) [13], the expected zone based on the location of a destination node is defined to reduce packet flooding. A number of redundant packets is dropped by limiting the flooding zone.

In addition, location information contributes to improving energy efficiency. In geographical adaptive fidelity (GAF) [14], nodes can go to sleep to conserve energy. Because GAF defines the virtual grid based on location information to find the nodes necessary for data delivery, it can maintain data delivery over on extended network life time.

Another application of location information is the MAC protocol for collision avoidance. Location information enables nodes to know the direction of neighbor nodes. In this case, the wireless communication range can be shrunk to avoid media access collision. In [15], the authors proposed that a node sends packets by using a limited wireless range with a directional antenna to a receiver when the transmitter knows the receiver's location.

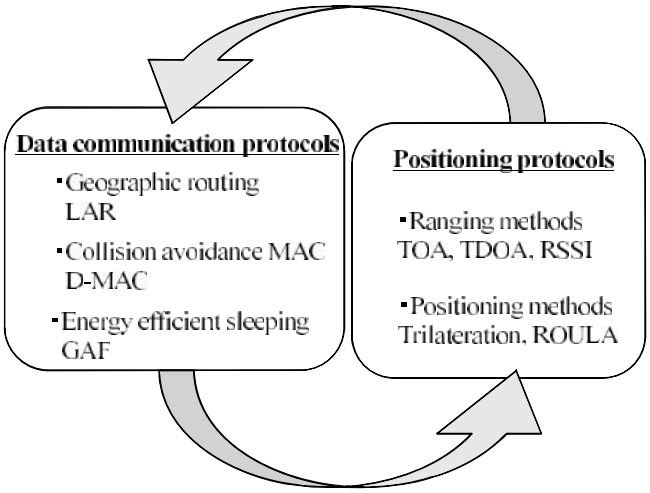


Figure 1: Relationship between data communication and positioning protocols.

While location information is useful for data communication protocols, a positioning protocol that provides node positions must be able to obtain location information at any place. As discussed in much literature, the global positioning system (GPS) is a simple solution to obtain node positions. However, GPS cannot always provide location information, such as for inside buildings.

A positioning protocol consists of two steps. First is estimating the distance (or ranging) by means of such as time-of-arrival (TOA), time difference of arrival (TDOA), and received signal strength (RSS). Second is positioning nodes to calculate the coordinates. Positioning methods for wireless multi-hop networks have been discussed previously [16], [17].

To enable nodes to obtain their positions at any place, a positioning protocol is required. The positioning protocol itself requires a ranging method that includes communicating nodes. Therefore, designing each data communication protocol and positioning protocol separately is not efficient. Hence, an integrated data communication and positioning protocol is designed.

1.1 Outline of paper

Section 2 describes an overview of the integrated protocol for communications and positioning. The physical layer (PHY) and medium access control (MAC) layer are described

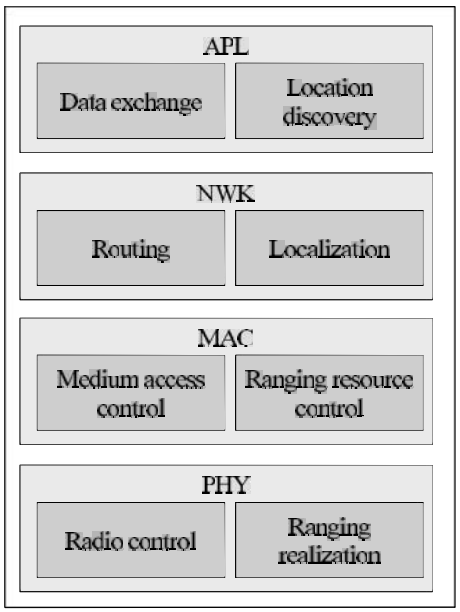


Figure 2: Protocol stack of integrated protocol for communications and positioning.

in Section 2.2 and 2.3, respectively. Section 2.4 presents an overview of the network (NWK) layer, including an issue of the layer and our solution.

A target tracking that is the application of the integrated protocol is introduced in Section 3. A cooperative target tracking using a ranging capability is proposed and the impact on positioning accuracy is investigated.

Section 4 summarizes the paper and mentions future work.

2 Integrated protocol for communications and positioning

2.1 Overview

The integrated protocol for communications and positioning is operated under a topology of wireless multi-hop networks.

The protocol stack of the integrated protocol is shown in Fig. 2. The objective of the integrated protocol is to enable simultaneous data exchange and location discovery. It enables users to obtain data and location information in the application layer (APL).

To enable data communications, each layer has conventional communications facilities, and to enable location discovery support, ranging realization, ranging resource control, and localization are added. We then describe the PHY, MAC, and NWK layers.

2.2 PHY

Ranging realization is achieved in the PHY. IEEE 802.15.4a [9] standardized the ranging capabilities, which enable nodes to estimate node distance, although the implementation is optional. We then use the same mechanism of ranging capabilities as described in IEEE 802.15.4a specification.

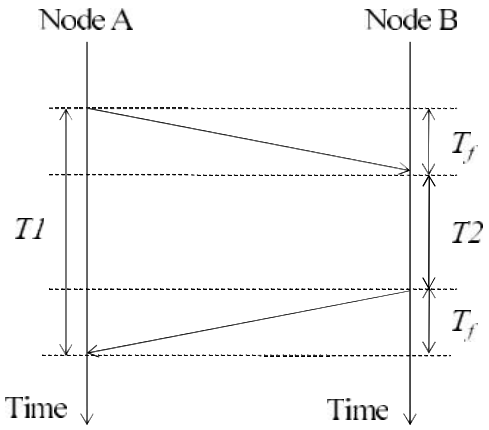


Figure 3: Ranging message sequence for TW-TOA.

Figure 3 shows the message sequence for ranging realization by using a two-way time-of-arrival (TW-TOA) between two nodes A and B. For our purpose, we only consider the true times of message arrivals. The propagation time T_f to estimate the node distance can then be written as

$$T_f = \frac{1}{2}(T1 - T2), \tag{1}$$

where $T1$ is the round trip time for node A and $T2$ is the reply time for node B. A distance can be derived from a propagation time, TW-TOA enables the estimation of node distance. TW-TOA can be achieved by observing the first arrival signal of received signals, and it requires at least two messages between nodes.

One observation about the IEEE 802.15.4a specification is that it enables the node to estimate distances in the PHY. However, IEEE 802.15.4a does not support a positioning method. Therefore, a facility to calculate node position is required to be defined on an upper layer.

2.3 MAC layer

Figure 4 shows the MAC protocol for the integrated protocol. Ranging capability enables tracking and navigation applications. In such applications, data may be required to be sent without collisions. The functionality in the MAC layer is to schedule and control the ranging resource to conduct the tracking and navigation applications. The contention access period (CAP) is collision-based packet scheduling, and the contention free period (CFP) is collision-free packet scheduling, which is the same mechanism as described in IEEE 802.15.4 [8] specification. Although CFP is required to the synchronization between nodes, data can be sent without collision. The CFP provide a guaranteed time slot (GTS) to send packets periodically. Thus, CFP can be used for the tracking and navigation applications.

A challenging functionality for the MAC protocol is how to select the resource to obtain the precise positioning accuracy. Positioning accuracy is the performance parameter to indicate the quality of service (QoS) in the integrated protocol. Each resource in the MAC layer is selected and scheduled for the

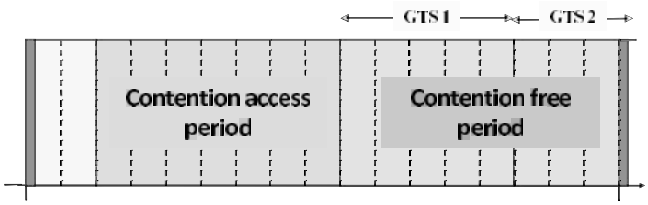


Figure 4: MAC protocol for integrated protocol.

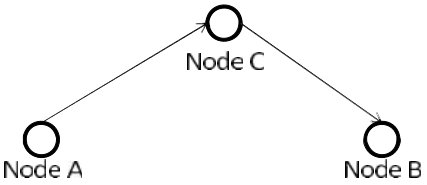


Figure 5: Example cooperative relaying for improving reachability.

precise positioning accuracy. Then nodes will cooperate for precise positioning accuracy. In existing data communication protocol, delay and reachability are parameters to indicate the QoS. Then, the nodes will cooperate for improve the delay or reachability. Figure 5 shows the simple case of cooperative relaying for improving the reachability. When reachability is critical for the network, node C will relay the data for node A and B. Because our interest is how node cooperate for precise positioning accuracy, we will investigate the performance of cooperative target tracking in Section 3. The simulation result in Section 3 will reveal that the cooperative relaying though a line-of-sight (LOS) link improves the positioning accuracy.

2.4 NWK layer

The NWK layer provides the simultaneous routing and localization capability. The localization protocol estimates node positions. The motivation for developing localization protocol is wanting to know the node positions in multi-hop networks with only a small number of anchor nodes. An anchor node is one whose position is known in advance through such as GPS.

We previously developed optimized link state routing-based localization (ROULA) [17]. ROULA is independent of anchor nodes and can determine the correct node positions in a non-convex network topology. In addition, ROULA is compatible with the optimized link state routing (OLSR) protocol [21] and uses the inherent distance characteristic of multipoint relay (MPR) nodes.

Our objective in developing the integrated protocol is to achieve simultaneous data exchange and location detection. The localization protocol consists of estimating distances and positioning. ROULA sends hello messages to estimate node distances, hence it can extract an overhead of routing protocol generated in the NWK layer.

Before discussing which routing protocol is suitable for the NWK layer, we first introduce existing routing protocols. The routing protocol is one of the major issues in wireless multihop networks. In the Internet Engineering Task Force (IETF),

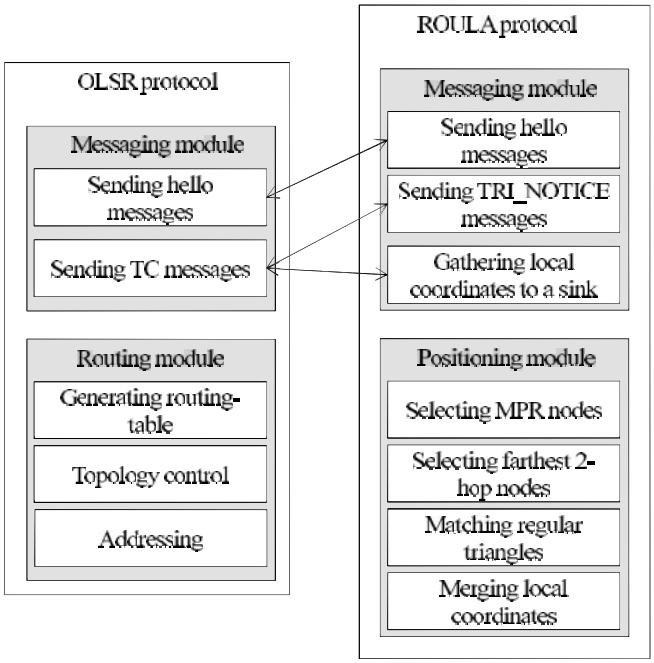


Figure 6: NWK protocols for integrated protocol. Arrows show compatibilities between OLSR and ROULA protocols.

Mobile Ad-hoc Networks (MANET) Working Group [2] has been organized to address this issue.

There are mainly two types of routing protocols. One is a reactive protocol and the other is a proactive protocol. Ad hoc on-demand distance vector routing (AODV) [20] is one of the reactive protocols. In AODV, control messages to detect and maintain the routes are generated according to requests. Zigbee [1], which is an sensor network product uses the AODV protocol.

OLSR [21] is one of the proactive protocols. OLSR sends the control messages periodically to detect the shortest paths to nodes in the network. Nodes in OLSR select the MPR nodes as relay nodes. OLSR enables efficient flooding of messages by using MPR nodes.

Figure 6 shows both the OLSR and ROULA protocols. OLSR and ROULA exchange hello messages to find one-hop nodes. In addition, OLSR uses topology control (TC) messages to find the routes in the overall network. TC messages periodically flood the network, hence they are compatible with the messages gathering local coordinates from all nodes in the ROULA protocol.

We selected the OLSR protocol for NWK layer. Currently, we are porting the ROULA protocol into the OLSR protocol. We are investigating how ROULA can be efficiently integrated in the OLSR protocol [18], [19].

3 Target tracking application

Target tracking is an application of the integrated protocol. We focus on investigating a target tracking using a ranging capability and describe how nodes on the integrated protocol operate for target tracking.

One of the applied situations of target tracking is inside

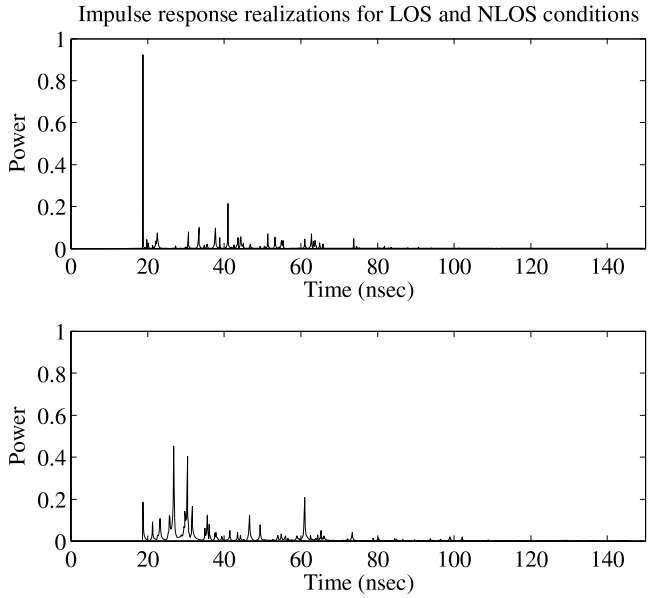


Figure 7: Time of arrival signals through LOS (top) and NLOS links (bottom).

a hospital. Target tracking enables the position of tags with ranging capabilities to be monitored. Hence, the positions of patients and doctors equipped with the tags can be known at once even in the case of a emergency situation.

The remainder of Section 3 is organized as follows. A problem statement is given in Section 3.1. Cooperative target tracking is proposed in Section 3.2. A performance evaluation by simulation is described in Section 3.3.

3.1 Problem statement

The problem of target tracking is how to estimate node positions sequentially. Although the problem to be solved is in a mobile node environment, we state the problem as a static location estimation for brevity.

Let us consider a two-dimensional positioning problem. Assume that at any time, the positions (x_i, y_i) for $i = 1...k$ reference nodes are known and the positions (x_i, y_i) for $i = k+1...n$ nodes are unknown. The typical location estimation using least-square is given by

$$\hat{p} = \arg\min \sum_{i=1}^{k} (r_i - \sqrt{(x_i - x)^2 + (y_i - y)^2})^2, \quad \text{for } k \geq 3, \tag{2}$$

where $\hat{p}$ is the estimated node position and k is the number of reference nodes. The range measurements obtained from TOA estimation is r_i. Equation (2) gives a good solution for the estimated position when the range measurements are done through LOS links. However, once the range measurements are done through non-line-of-sight (NLOS) links, the estimated position will be biased.

Let us consider that the range measurements are

$$\hat{r_i} = r_i + \begin{cases} e_i^{los}, & i = 1, 2, ..., M \\ e_i^{los} + b_i^{nlos}, & i = M+1, ..., N \end{cases}, \quad \text{where } e_i^{los} \sim \mathcal{N}(0, \sigma^2),\ b_i^{nlos} \sim \mathcal{E}(\mu). \tag{3}$$

LOS measurement noise e_i^{los} is modeled as a zero mean Gaussian distribution with variance σ^2. NLOS bias b_i is a positive distance bias introduced due to LOS blockage, and is modeled as an exponentially distributed random variable with mean μ.

Figure 7 illustrates typical impulse response realizations for ranging at the receiver for LOS and NLOS conditions. In the case of a LOS link, the first arrival signal is normally identical to the shortest path signal of the sight. However, when an obstruction blocks the LOS link, first arrival signal received at the receiver may not be identical to the signal of the shortest path. Reflections from scatters are reached at receiving nodes. Hence the range measurement through NLOS link results in introducing bias error.

3.2 Cooperative target tracking

3.2.1 Notations and assumptions

Let us introduce notations for the three devices that we used in the proposal as listed in Table 1. A target node (TN) is a node that should be tracked and estimated its position. A mobile node (MN) is a node that has the capability to move and has a role in assisting the TN tracking. An MN can be a human or a mobile robot. Reference nodes ($\mathrm{RN_i}| \ i = 1 \ldots 3$) are the nodes whose positions are known. Additionally, we make the following assumptions.

- TN, MN, and RN have TOA ranging capability.
- Identifications for LOS/NLOS links are achieved by using simple hypothesis testing of received signals in a mobile node environment [7].
- NLOS link is generated when an obstruction crosses a LOS link, for simplicity as illustrated in Fig. 9.

Figure 9(a) presents the guaranteed area to obtain LOS links from three RNs. When the MN is placed in the shaded areas, it is guaranteed to obtain LOS links from three RNs. Figure 9(b) presents the guaranteed area to obtain LOS links from three RNs and TN.

3.2.2 Overview

Introducing the positioning accuracy as a parameter of QoS in the integrated protocol motivates that nodes in the network cooperate for precise positioning. A cooperative target tracking using mobile nodes is then proposed to obtain precise positioning in an NLOS environment.

For brevity, we describe the proposal by assuming a static snapshot as illustrated in Fig. 8. Three RNs, an MN, and TN are located on a field. The problem is to estimate the TN

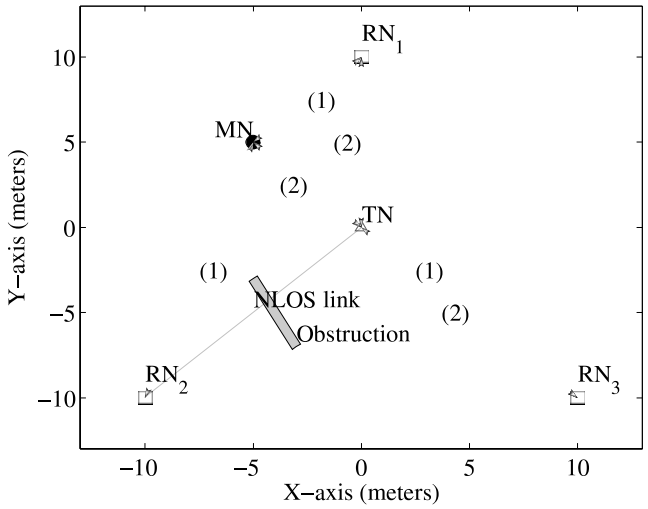


Figure 8: Illustration of target tracking. Messages are indexed in number order for cooperative target tracking.

position. An obstruction is located between TN and RN_2, resulting in that estimated position having bias due to the range measurement through NLOS links. To avoid such a positioning situation that includes NLOS links, MN moves to an area to obtain LOS links from the TN and RNs when the TN has an NLOS link for positioning.

In Fig. 8, MN is located to have LOS links from the three RNs and TN. We describe an area that is guaranteed to obtain LOS links in Section 3.2.3. In this case, the RNs estimate the TN position by relaying the MN positioning. First, the RNs estimate the MN position with Equation (2) through LOS links of indexes (1) as described in Fig. 8. Once MN obtains its own position, it can be considered as a pseudo reference node. Then, RN_1, RN_3, and MN estimate the TN position with Equation (2) through LOS links of indexes (2) as described in Fig. 8.

3.2.3 Guaranteed area to obtain LOS links

To assist TN positioning, MN moves to the guaranteed area to obtain LOS links from three RNs and the TN.

At present, we only consider a situation where an area to obtain LOS links from three RNs and TN always exists, and MN is placed that area.

To cope with general cases, we are investigating the following point:

- The condition for a guaranteed area where LOS links can be obtained from k RNs and TN (x, y), where $k \geq 3$, and x and y are variable.

3.3 Performance evaluation

3.4 Simulation setting

A performance evaluation through simulation showed the effectiveness of the proposal. We implemented extended Kalman filtering (EKF) [5] to estimate moving TN positions. A node's motion can be considered as a dynamic system for a function of time. Therefore, EKF can be applied to target tracking [5], [6].

Table 2: Notations for three methods.

Notation	Description
Conv.	Non-cooperative target tracking using EKF
Proposal	Cooperative target tracking using EKF

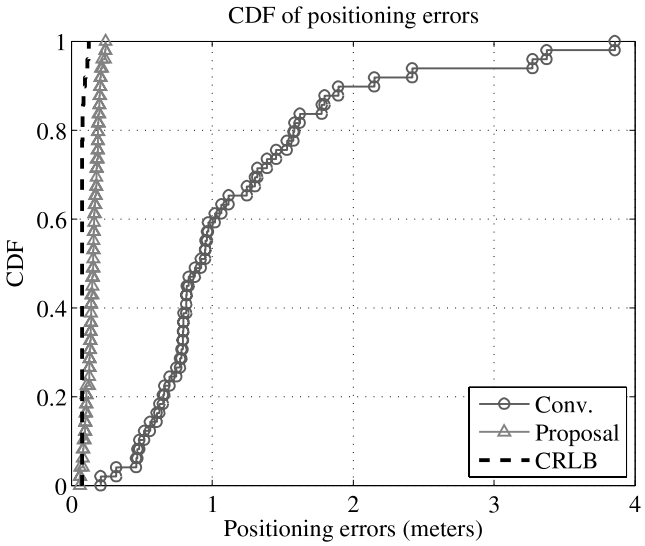


Figure 11: CDF of positioning errors.

Table 2 lists notations for two methods whose performances was compared. Conventional method represents non-cooperative target tracking using EKF that estimate moving TN through NLOS links.

In addition, we implemented Cramer-Rao lower bound (CRLB). CRLB is the bound on a unbiased estimator [10]. This bound provides the best achievable performance [5]. Therefore, it can be used as a performance benchmark of the accuracy of target tracking. The covariance matrix of EKF in the absence of a process noise equation is identical to CRLB [11].

3.5 Simulation result

The cooperative target tracking was performed by using a simulation. For LOS range measurement and NLOS bias error, we used $\sigma^2 = 0.1$ and $\mu = 1.6$ (m).

Figure 10 illustrates the estimated positions for the conventional method (left) and the proposed method (right). RNs are placed with $\mathrm{x}_1 = [0, 10]$, $\mathrm{x}_2 = [-10, -10]$, and $\mathrm{x}_3 = [10, -10]$ (m). The start point of TN is at $[0, 0]$ (m). TN moves in the straight line at the constant velocity of 1.0 (m/s). The observed time is 5 (s) with sampling interval of 0.1 (s). A link between TN and RN_2 is blocked by the obstruction.

As shown in Fig. 10, estimated positions by conventional method had biased positioning errors from RN_2. We observed that the estimated positions by the proposed method were much closer to actual TN trajectories.

Figure 11 plots the cumulative distribution function (CDF)

Table 1: Notations for three devices.

Notation	Description	Object
Target node (TN)	Position should be tracked.	Human
Mobile node (MN)	Relay for TN positioning through LOS links from TN and RNs.	Human or mobile robot
$\mathrm{RN_i}\mid i = 1 \ldots 3$	Position is known.	

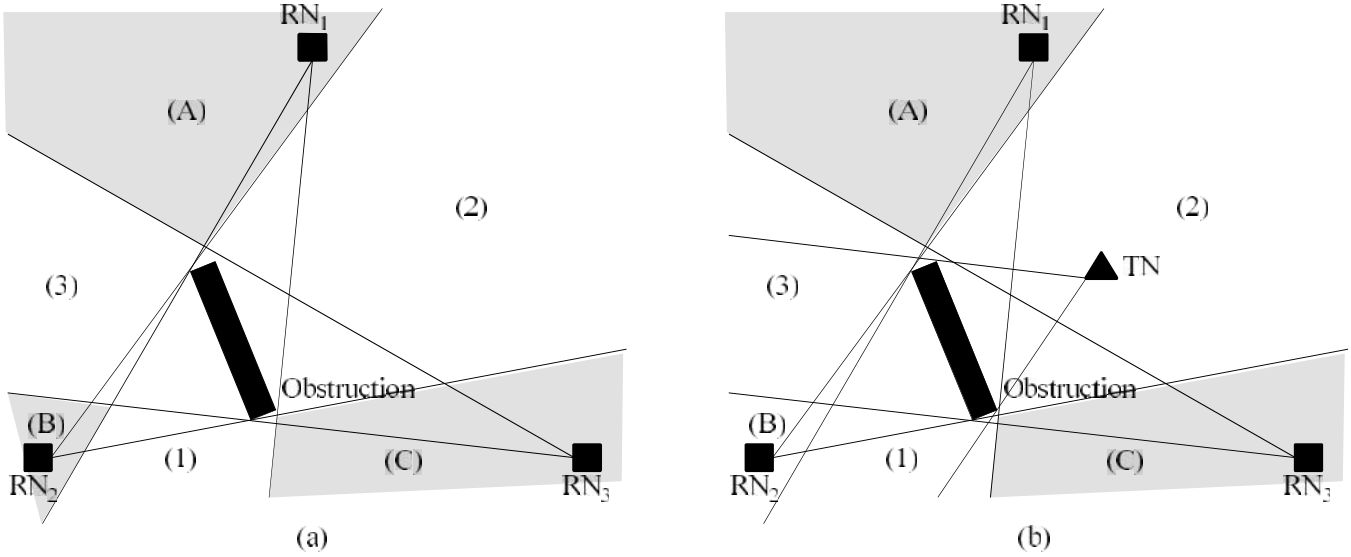


Figure 9: Illustration of areas (shaded) that are guaranteed to obtain LOS links from (a) three RNs and (b) three RNs and TN.

of positioning errors that are defined as

$$\sqrt{(x_t - x_t^A)^2 + (y_t - y_t^A)^2}, \tag{4}$$

where (x_t, y_t) denotes the estimated position at time t and (x_t^A, y_t^A) denotes the actual position at time t. Conventional method had large positioning errors. The proposed method was less positioning error and approached the CRLB.

We only presented the one scenario. Various scenarios including several obstructions and random target motions will be investigated in the future. Besides the positioning accuracy, delays in using the time slot for positioning in the MAC and collecting the data will be considered.

4 Summary

We presented an overview of an integrated protocol for communications and positioning. The objective of this integrated protocol is to enable simultaneous data exchange and location discovery, simultaneously. We described each layer of the integrated protocol. For the MAC layer, the resource control for positioning was introduced. For the NWK layer, we described compatibility of OLSR with the localization protocol for enabling to simultaneous routing and localization. We also investigated cooperative target tracking. Using a simulation, we found that cooperative target tracking achieved less positioning errors and approached the CRLB.

In future work, we will conduct detailed evaluations of node cooperation for precise positioning and simultaneous routing and localization.

REFERENCES

[1] Zigbee Alliance, http://www.zigbee.org/.

[2] MANET Working Group, http://www.ietf.org/html.charters/manet-charter.html.

[3] J. J. Caffery and G. L. Stuber, "Overview of Radiolocation in CDMA Cellular Systems," *IEEE Communications Magazine*, vol. 36, pp. 38–45, 1998.

[4] N. Patwari, J. Ash, S. Kyperountas, A. O. Hero, R. M. Moses, and N. S. Correal, "Locating the Nodes: Cooperative Localization in Wireless Sensor Networks," *IEEE Signal Processing Magazine*, vol. 22, pp. 54–69, 2005.

[5] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, "Estimation with Applications to Tracking and Navigation," *Wiley-Interscience*, 2001.

[6] A. Smith, H. Balakrishnan, M. Goraczko, and N. Priyantha, "Tracking Moving Devices with the Cricket Location System," *Proc. of ACM International Conference On Mobile Systems, Applications And Services (Mobisys)*, pp. 190–202, 2004.

[7] J. Borris, P. Hatrack, and N. B. Manclayam, "Decision Theoretic Framework for NLOS Identification," *Proc.*

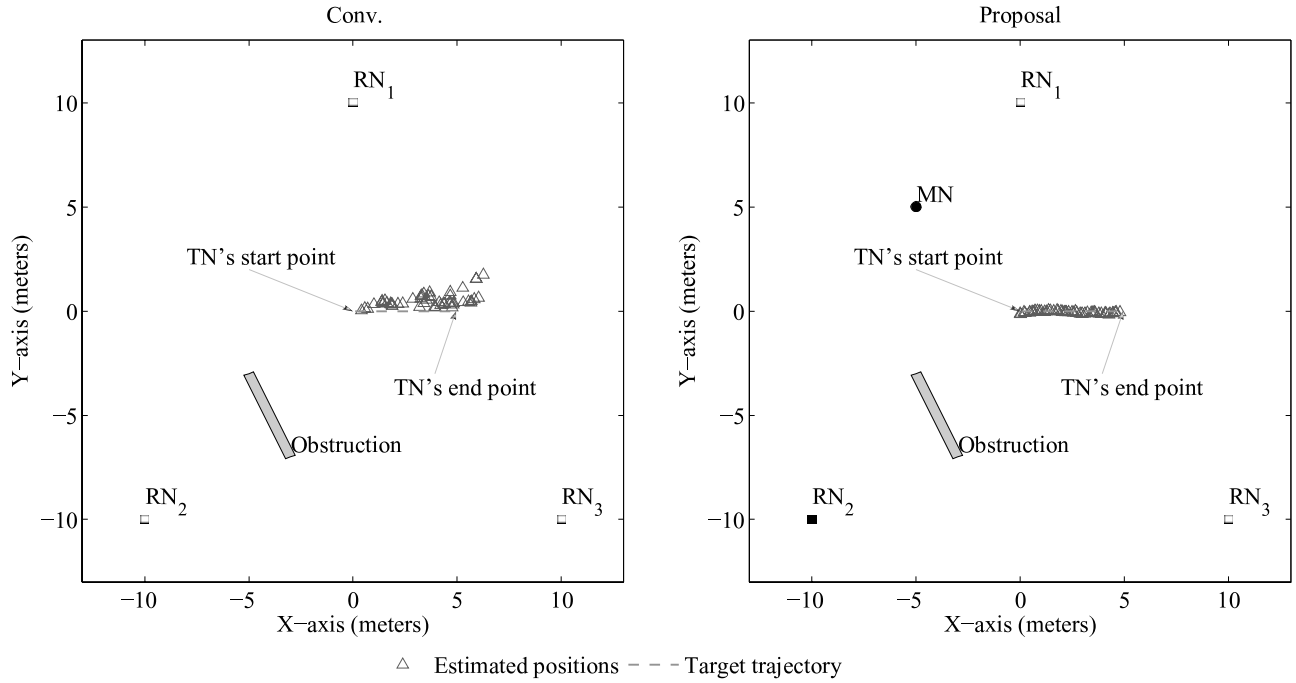


Figure 10: Estimated positions for conventional (left) and proposed (right) tracking methods.

of Vehicular Technology Conference (VTC), vol. 2, pp. 1583–1587, 1998.

[8] IEEE Std 802.15.4, "Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (WPANs)," *IEEE Computer Society*, 2006.

[9] IEEE Std 802.15.4a, "Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (WPANs), Amendment 1: Add Alternate PHYs," *IEEE Computer Society*, 2007.

[10] H. L. V. Trees, "Detection, Estimation and Modulation Theory, Part 1," *Wiley-Interscience*, 2001.

[11] J. H. Taylor, "The Cramer-Rao Estimation Error Lower Bound Computation for Deterministic Nonlinear Systems," *IEEE Transactions on Automatic Control*, vol. 24, pp. 343–344, 1979.

[12] J. Navas and T. Imielinski, "GeoCast – Geographic Addressing and Routing," *Proc. of ACM/IEEE Mobicom*, pp. 66–76, 1997.

[13] Y. Ko and N. H. Vaidya, "Location-aided Routing (LAR) in Mobile Ad Hoc Networks," *Proc. of ACM/IEEE Mobicom*, pp. 66–75, 1998.

[14] Y. Xu, J. Heidemann, and D. Estrin, "Geography-informed Energy Conservation for Ad Hoc Routing," *Proc. of ACM Mobicom*, pp. 70–84, 2001.

[15] Y. Ko, V. Shankarkumar, and N. H. Vaidya, "Medium Access Control Protocols Using Directional Antennas in Ad Hoc Networks," *Proc. of IEEE Infocom*, vol. 1, pp. 13–21, 2000.

[16] A. Savvides, C. Han, and M. B. Strivastava, "Dynamic Fine-grained Localization in Ad-hoc Networks of Sensors," *Proc. of ACM/IEEE Mobicom*, pp. 166–179, 2001.

[17] T. Takenaka, H. Mineno, Y. Tokunaga, N. Miyauchi, and T. Mizuno, "Performance Analysis of Optimized Link State Routing-based Localization," *Journal of Information Processing Society of Japan (IPSJ)*, vol. 48, no. 9, pp. 3286–3299, 2007.

[18] K. Soga, T. Takenaka, Y. Terashima, Y. Tokunaga, H. Mineno, and T. Mizuno, "Proposal of the Fusion Technique of Localization and the Routing Protocol," *Proc. of IPSJ National Convention*, pp. 187–188, 2008 (in Japanese).

[19] K. Soga, T. Takenaka, H. Mineno, Y. Tokunaga, Y. Terashima, and T. Mizuno, "Proposal and Evaluation of Uniting Technique of Localization and Routing," *Proc. of Multimedia, Distributed, Cooperative and Mobile (DiCOMO) Symposium*, pp. 120–126, 2008 (in Japanese).

[20] C. Perkins, E. Belding-Royer, and S. Das, "Ad Hoc On-Demand Distance Vector Routing," *IETF RFC 3561*, 2003.

[21] T. Clausena and P. Jacquet, "Optimized Link State Routing Protocol (OLSR)," *IETF RFC 3626*, 2003.

Reliable Method for Collecting and Evaluating Transmission Records for MANET Forensics

Akira Otaka†, Osamu Takahashi‡, and Tsuyoshi Takagi‡

†Graduate School of Future University - Hakodate, System Information Science
‡Future University - Hakodate, Faculty of System Information Science
{g2108008, osamu, takagi}@fun.ac.jp

Abstract - In recent years, mobile ad-hoc networks (MANETs) have become more popular; they can be autonomously constructed by the mobile devices without any communication infrastructure. However, it is reported that there have been several attacks on MANET networks. Furthermore, several intrusion detection systems (IDSs) for MANET are being discussed. However, as the IDS is not a perfect system, it may mistake a normal node for a malicious one. It is important that a node can prove accurate information on the communication status (ex. relaying packets) to a third person. In other words, studying MANET forensics is essential.

We first define the MANET forensics model, and propose which contents should be collected as evidence of an attack, and the method of retrieving evidence packets from the witness node. Furthermore, we implemented a prototype system on the basis of the proposed method, and evaluated the performance, (the influence on the network traffic, amount of the collectable evidence, etc.) using the network simulator. Finally, we conclude that our proposal is suitable for MANET forensics and discuss the subject of future research.

Keywords: Wireless Networks, Mobile Ad-hoc Networks, Network Forensics, MANET Forensics

1 Introduction

In recent years, wireless networks researchers have studied new wireless networks that do not depend on existing infrastructures. In particular, a MANET can be constructed with the participants own resources is gaining a great deal of attention. The advantage of this network is that it is inexpensive to start and maintain because it is constructed with mobile devices. An important application of MANETs is providing communication at the site of a natural disaster and people who work in these sites expect MANETs to continue to be usable even in unfavorable conditions, which may include being at sea or in the air. However, as the number of ways that a MANET has been identified, many ways to defend it have been proposed.

2 Related work

A flow model used to collect evidences in a MANET is shown in Figure 1. A relay request creating evidence that the data is sent by the user to witnesses. The witness who received the packets requested evidences creates and sends it to the relay. The witnesses create the evidences as follows:

1. Hashes with the data packets of MANET forensics packet payload (include ethernet payload).
2. Encrypts the value hashed(1) with their secret key.
3. Stores the value hashed by them (1) and the code encrypted by them (2) to a payload of the MANET forensics packet.
   - Type of MANET forensics header is changed.
4. Witness sends the MANET forensics packets.

We concluded that, since the evidence satisfies the requirements of RFC3227[1], the evidences are reliable and fair. However,

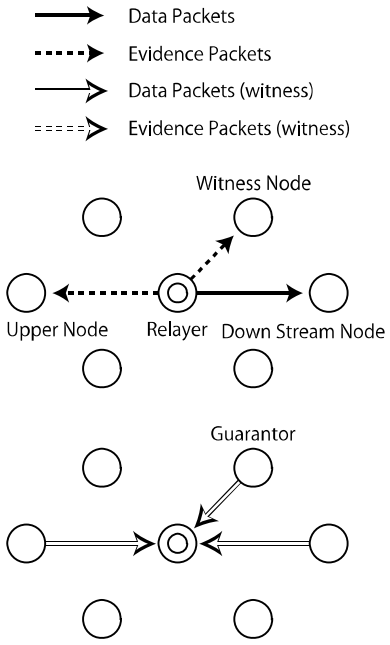


Figure 1: Methods of Collecting Evidences

evidence does not necessarily gain enough reliability to be used in a court case.

3 Proposed Method

3.1 Impetus for MANET forensics

A suspect node detected by a detecting node is immediately reported to a source node that then starts the function of

gathering evidence.

3.2 Methods of Collecting Evidence

We defined E_{part} is number of evidences collected by a relayer, E_{all} is number of evidences collected by relayers, and F is number of relayers. Expression is showed E_{all} as a follow expression:

$$E_{all} = \sum_{f=0}^{F+1} E_f \quad (1)$$

$F + 1$ includes relayers and a source node. We adopt some one of follow systems to collect evidences with environment.

(1) Broadcast Evidences System

For networks with low reliability, a relay request that data is sent by the potential user for all witnesses. A great deal of evidence collected by the relay, but doing this means that the networks are loaded heavily.

(2) Particular Evidences System

A relay requests provides evidence for high-integrity nodes when it relay knows what about all the nodes around the central nodes (the other relays). This system reduces the evidence packets more effectively than the Broadcast Evidences System does. However, there is a fault: if no one knows what about the surrounding nodes, the system is inoperative.

(3) Fast Evidences System

If a relay required evidence does not know the position of the nodes around it, then it collects evidence with the Broadcast Evidences System. After that, the relay recognizes surrounding nodes and collects evidence with a particular evidences system. This system has two advantages. First, the relay does not need to know the what of the surrounding nodes beforehand. Second, the networks place a low load because of the decrease in the number of packets created by the broadcast evidences system. Thus, this system is highly effective and efficient.

3.3 Contents of the Evidence

When a relay is suspected of being an attacker, it bears the responsibility to prove that it has sent the data. To enable to the system to do this, we defined the contents of the evidence as follows.

- Contents of Sent Data
- Time of Creating Evidences
- Information relating to Guarantors
- Option
  - Geographic Information of Creating Evidences from Geographic Information System(GIS)

Witness nodes requested to create evidence of sent data from a node of collecting evidences create evidence packets. Thus, the victim preserves the evidence of entire packets that includes the ethernet headers used for chasing attackers. However, there is no sense of preserving the evidences of whole packets in the MANET because of the difficulty of giving chase relays. To control MANET forensics, we saved the ethernet payload without a header. A guarantor hashes the ethernet payload to get a small and specific amount of it and protect its privacy when it is analyzed. The evidence is reliable because it is impossible to doctor the evidence through changing the computing digital signature. We used $wNAF$[2] cryptography that was also proposed for mobile machines because, in general, RSA public-key cryptography takes too long to compute. However, we are able to change the public-key cryptography system if needed. A guarantor creates evidences and bears data sent by relays. We need a public key of the guarantor if we wish to validate the evidence. Thus, we need to save information of guarantors to get their public key. A time stamp of evidence preserved by relays and packets of evidences by guarantors is used to find discrepancies in the evidences. Furthermore, MANET can be used anywhere. Therefore, a GIS is more reliable if the following are used: geographical evidence that includes the GIS of evidences preserved by relays and evidence packets created by guarantors.

4 Prototype implementation and evaluation

We proposed a method of collecting evidences by using witness nodes[3]. Therefore, the method is detrimental to communication since a lot of packets conflict with others. We found that varying the number of evidence packets with changes in the number of surrounding witness nodes and influences the whole networks in the process of collecting evidences.

4.1 Prior Condition

4.1.1 Detecting Method

To enable incidents to be detected and thus enhance security, the network is treated with extended watchdog function proposed by Yokoyama et al.[4]. Doing this puts no load on the networks because the packets are not changed with each other when selfish-nodes are detected. We focused on on how to collect evidences after an incident was detected and propose using this method.

4.1.2 Encrypting Method

We postulate that using public-key cryptography when data packets are encrypted enables the participants in the network to create the evidence packets. Also, public-key paring with secret-key encrypting the evidence packets is available. Therefore, the evidence packets encrypted by the secret-key are always checkable.

4.2 Experimentation Environment

A main aim our experiment is to find how using MANET forensics affects the load placed on a network. Therefore, we did not focus on the attacking nodes and the function of detecting, and evaluated the cases of MANET forensics that are currently available. Since we used the premise that the most appropriate methods of collecting evidence packets are

the Broadcast Evidences System and the Fast Evidences System and all nodes move at 1.4 m/s in this experimentation, we gained data from the worst case pattern.

4.3 Simulator Environment

An environment using the simulator is shown in Table 1. We used the following: "Network Simulator-2[5], 802.11 in the MAC layer, AODV in the network one, UDP in the transport one, and CBR in the application one.

Table 1: Simulator Environment

Position Range of Nodes	1000 * 1000 [m]
Communication Range	250 [m]
Mobility	1.4 [m/s]
Simulation Period	100 [s]
Number of Nodes	10 - 100 nodes
Communication	Protocol: CBR/UDP/IP Occurrence Period: 0.05 [1/s] Size of a Packet: 1000 [Bytes]

We put number of nodes between 10 and 100 by 10 nodes move with 1.4 [m/s] in random topology per thousand square meter. A communication pair, a source node and a destination node, is selected in random order. Network traffics have been occured per 0.05 [1/s] for a hundred second. In this regarad, the atterckers do not exist.

We set the number of nodes between 10 and 100 by 10 nodes move with 1.4 m/s in random topology per thousand square meter. A communication pair, a source node, and a destination one, are selected in random order. Network traffic occurred per 0.05 [1/s] for a period of a hundred seconds. In this environment, the attackers do not exist.

4.4 Metrics for Evaluation

We use metrics as follows to evaluate those methods.

(1) Packet Delivery Rate

Packet delivery rate is defined by the goodput for each bit rate. Thus, to calculate the goodput, we divided the number of packets received by a destination node by the number of packets that were sent to the source node. The packet delivery rate is the highest the nearer the value of the good put gets to one.

(2) Number of All Packets

The number of all the packets is packets exchanged between the networks. This includes packets that are classified by controlling networks, data, and pieces of evidence. However, duplicative packets received by the same node are not counted. That is, the number of all packets shows the load on the network. The goodput goes down when this metric is increased; therefore, to ensure that communication channels stay stable, we need to keep the value of the number of all packets constant.

(3) Average of Evidences

This metric is the average number of pieces of evidences that maintained their integrity, and is measured by the number of relays per datum. The number of pieces of evidence shows the level of reliability. Thus, a high level of reliability pays on the delivery of becoming unstable networks.

4.5 Results and Discussion

Packet delivery rates per number of nodes and the methods are shown in Figure 2. In all the cases, the packet delivery rates are increased for 50 or fewer nodes. When the MANET function is used, the packets are then received by 95% or more of the destination nodes. However, the rate decreases with the function of the MANET forensics. In particular, the rates decreased by 40% less than the decrease when the Broadcast Evidences System was used without the MANET function. On the other hand, the rates of Fast Evidences System decreased by 20% less than when the function of the Fast Evidence System was used with the Broadcast Evidences System because the Fast Evidence System is made by improving the Broadcast Evidences System. Since the rates are increased for number of 50 nodes or less in all of the case, using MANET forensics does not have any effect when density is 50 [1km2] or less.

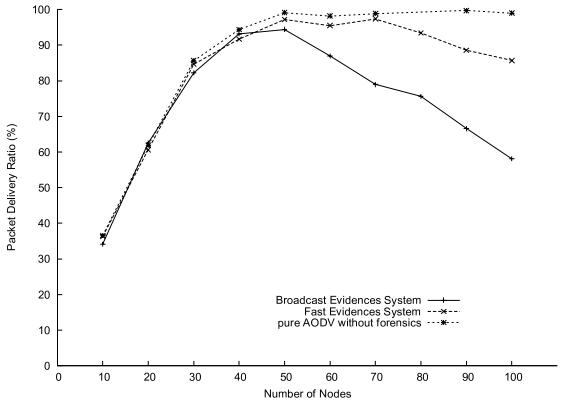


Figure 2: Packet Delivery Rate

The number of all packets is shown in Figure 3. A summation of the packets relayed and sent in the networks, i.e., the network load is shown. When no MANET forensics are used, the number of all packets steadily increased for 50 nodes or more. However, when the Broadcast Evidence System is used, the number of all packets greatly increased with an increasing number of nodes. Because the number of witness nodes increased with an increasing density of participants, more evidence packets are sent by the witness nodes. On the other hand, using the Fast Evidences System to restrict evidence packets produced a 50% increase in performance. Despite this, using MANET forensics increased the number of all packets more than when none were used.

The average of evidence packets received by a relay is shown in Figure 4. The Broadcast Evidences System places the burden on the networks since the number of evidence packets

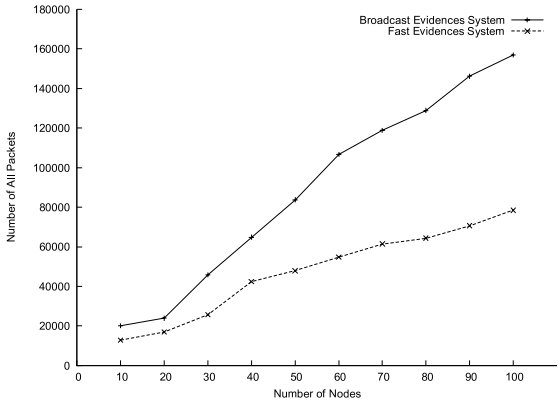


Figure 3: Number of All Packets in Difference Methods

increased by an increasing number of witness nodes.

Furthermore, the averages of evidence packets converge as shown in Equation (3) with the function of the Broadcast Evidences System.

$$E_{ave} = N_{all} * \frac{Communication\ Range}{Experimental\ Range} \quad (2)$$

$$E_{all} = (F+1) * E_{ave} \quad (3)$$

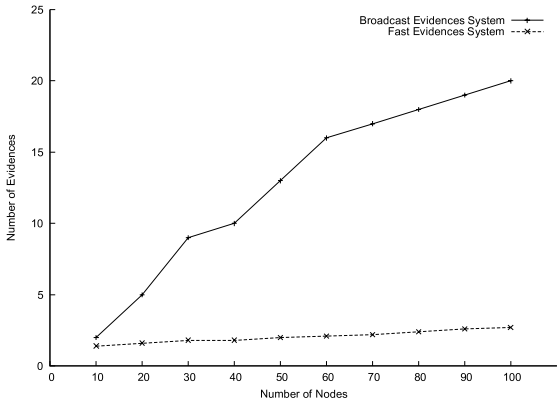


Figure 4: Average Number of Pieces of Evidence Received by Relay Using Difference Methods

5 Conclusion

We propose collecting evidence that is highly reliable. However, as there is a possibility that the process of collecting evidence significantly loads the networks, we need to check that doing this will not damage the other communications. We evaluated the packet delivery rates, the number of all packets, and the average of the number of pieces of evidences in model of collecting evidences by all relays. As a result, the proposed method showed that is a trade-off between the reliability and that rates of goodput and network loads are vary widely with selected methods. Furthermore, we confirmed that our process of collecting those evidences worked as we had designed it to. In this evaluation, the relays collect evidences effectively and we expect that the improvement of those performance will be higher than when using the Fast Evidences System. This is because relays of downstream from the detector node are used to collect evidence. However, we assumed that the function of MANET forensics starts from the source node.

6 Future Work

6.1 Policy of Networks (Timing when to start)

The number of packets of evidence is significantly changed with the timing of the starting function. For example, relays that are downstream from the detector node immediately collect evidence. However, such evidence has a low level of reliability because of there is no way to draw a comparison between it and the evidence collected from the upper nodes. Thus, to ensure that evidence is collected in a fair and impartial way, our policy was to apply the forensic technique to whole networks.

6.2 Consideration of Evidence Packets

We focused on collecting evidences. However, network forensics involves completing a cycle of applying policy, collecting, and analyzing evidence. Thus, we need to analyze the evidence and give feedback to the network. We can impose a sanction, i.e., reject the communication, and assign the network resources to nodes made a contribution to relay to the other packets by priority in the whole networks. We expected that doing this will decrease the risk that the system will be attacked.

REFERENCES

[1] D. Brezinski and T. Killalea, Guidelines for Evidence Collection and Archiving, RFC 3227 (Best Current Practice) (2002).

[2] K. Okeya and T. Takagi, The width-w naf method provides small memory and fast elliptic scalar multiplications secure against side channel attacks, IEICE Transactions, Vol.E87-A, No.1, pp.75–84 (2004).

[3] A. Otaka and O. Takahashi, Network forensics on adhoc networks, IPSJ SIG Technical Reports, Vol.2008, No.18, pp.217–222 (2008).

[4] S. Yokoyama, Y. Nakane, O. Takahashi and E. Miyamoto, A proposal of security system with high-precision detection of misbehaving nodes in ad hoc network and its evaluation of implementation, Transactions of Information Processing Society of Japan, Vol.49, No.2, pp.639–649 (2008).

[5] nsnam, The network simulator - ns-2, `http://www.isi.edu/nsnam/ns` (2008).

Session 2: Business System (Chair Hisao KOIZUMI)

Structured Information Design Methodology for quality decision in competitive business arena

Kazuya Matsudaira*, Teruhisa Ichikawa** , and Tadanori Mizuno*

*Graduate School of Science and Technology, Shizuoka University, Japan
{matsudaira, mizuno}@mizulab.net
**Faculty of Informatics, Shizuoka University, Japan
ichikawa@inf.shizuoka.ac.jp

Abstract -This paper describes Information Design way for helping people get high quality of information in the enterprise. Information is deemed as vital asset at enterprise, but in a reality, most enterprises are not confident they can bring order out of information chaos. Because of chaotic environment, too much work is repeated and people are not sharing information. By restructuring information along the predetermined four level of structure, we are able to share and reuse the important resources to reduce cost and cycle time. To arrive at Information Resource Managed situation, we suggest that enterprise must design information by top-down approach. With this approach, at all levels of organization, both managements and employees may get very adequate information to make decisions or actions. Enterprise will be successful in orchestrating market changes with self defending improvements of organization in this uncertain age.

Keywords: Information Structure, IRM, Decision-making, Methodology

1 INTRODUCTION

Enterprise will not only survive, or sustain but prosper in the 21st century.

Success of enterprise depends upon quality of information. That is to say, quality of decision making is the driving force for success. However, information infrastructure is hardly vulnerable. In the past century, we never thought that information is asset.

1) Accuracy of data is low.
2) Reliability of manually created information is doubted.
3) Timing of information is inadequate.
4) Work load fluctuation causes increase of clerical workforces.
5) Information requests by Government, Unions and competitive pressures are neglected.

To improve situations above, enterprise must redesign information taking the value of it into consideration. Otherwise, enterprise can not keep growing with sustainability.

If we design information required for each level of organization, those quality information will be useful for people who wants to decide and act resulted in profitable performances [1].

2 UNDERSTANDING OF INFORMATION

Enterprises need information for performing works. Top management requires strategic information for decision. At second level division directors need tactic information for them to decide. Managers or leaders would use administrative information to manage his/her section or team. Even, operators at factory, or clerks at office may work, triggered by information to produce products and services. Information they need at each hierarchal level are a variety of kinds.

We made detail analysis and found that decision maker at each level is dependent upon information types at X company that is a manufacturing/marketing company of Electronic/Communication instruments. Regarding Information, we define that Information is not equal to Data. Information is produced from Data. Information is consumable commodity. It is used by people. People only use information, and computers process data and produce information. Computers never process Information. Information supports the actions and decisions of the enterprise. Such actions and decisions must be done within a unique time-frame [2].

Therefore, it is time dependent. It is common that we never store information.

If we store data, then we can retrieve data and may produce information within the timeframe. Organization could progress when the good impact of both actions and decisions outweighs the impact of bad actions and decisions.

Information is designed on OUTPUT as a media which conveys information to people. Data is collected by INPUT that inputs data to FILE and data is stored in FILE, classified in a RECORD.

Data is assorted into RECORD in the FILE. DATA-BASE is a conceptual set of integrations of FILES.

Information system is designed by information driven way. This means that design process starts from requirements definition. Information design never start

from hardware or media selection. Information driven design principle is to be kept.

3 FOUR LEVELS OF INFORMATION STRUCTURE

Figure 1 shows 4 levels of Information structures, also it shows the structures of SYSTEM / ORGANIZATION / DATABASE that are 4 aspects of Information Resources. Top level of Information is called Strategic information. The second level is named Tactic information. The third level is said Administrative information for policy communication and controlling performance. The fourth level is defined as Operational information as a bottom [3][4].

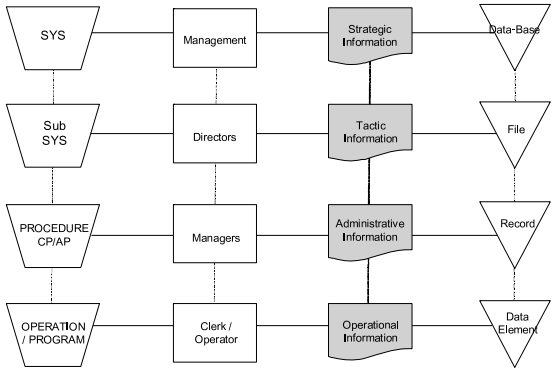


Figure 1: 4 Aspects & 4 Levels of Structure

This figure teaches that Enterprise managements are likely to use Strategic Information via some Data-Base through the System for making decisions.

Division Directors are basicly depending upon the Tactical Information for maintaining or improving the performances of some divisions or group-company.

In this case, they use multiple sub-systems to get Tactic information via some Files.

Managers need Administrative Information to control Section/groups manual procedures or computer procedures. The information is derived from some Records in the File.

Operational Information is used by the bottom level of organization structure, operators in the factory or clerks in the office. They are usually handle some sort of Data-Elements relating to 5W-1H. In the modern enterprise, computer displays are used to process Data-Elements quickly. Operators/clerks keep routine work by predetermined manual procedures/computer programs. Through those works, many Data-Elements are up-dated or maintained every day.

Figure 2 is an example of structured information designed by us and used at X company.

In the information at bottom level, you may see data element but those data elements are processed into generated data through formula and appears in the information at upper level. It is rare that you may find some data elements in the Strategic information.

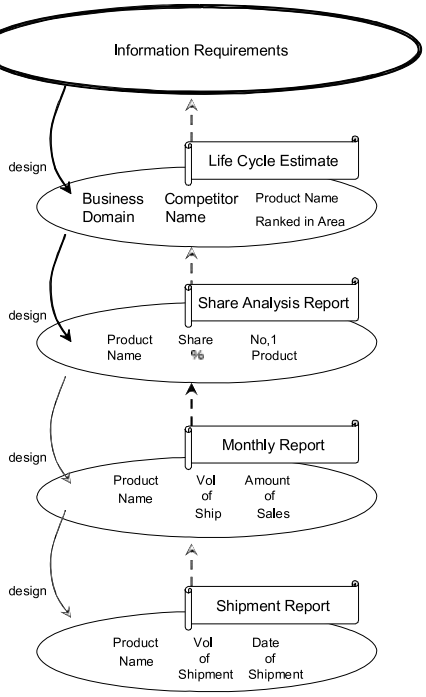


Figure 2: Example of Information Structure

4 ENGINEERING DESIGN PROCESS OF INFORMATION

10 phases are prepared as a methodology for the design and develop information from information requirements. At Phase 1, we plan and collect information requirements. Then we design Strategic information at top level. We should confirm by checking of formula all generated data into data elements. In this process, we may make a draft of directory of data. The typical ISD (Information Systems Department) still has no viable integrated repository. We developed a directory which has about 10thousand of data elements. Also, there were no accurate, coherent documentations of the valuable information resources at ISD. At this phase, we might build models of the repository for metadata contents. In order to manage resources, here, at phase 1, we should evaluate functionality of repository and select and build

it for future use of information on information. The way is to appeal to CIO about credibility of IS D [5][6][7].

At Phase 2, we design Tactic information for division managements to achieve targets objects of division. Phase 3 is a design phase of Administrative information that support managers who act as section head. Managers must pay to 360 degree orientations. Phase 4 is a final design phase for Operational information before physically implementing Output. At this phase, we pay careful attentions to works by operators/clerks. Phase 5 is a detail design and creation of all OUTPUT in the physical design phase. Phase 6 is a trial test phase to get correct Output.

Phase 7 is an education step for actual users to learn outputs and inputs and files. From Phase 8, we are ready to use Information for real jobs. Phase 9 is auditing phase, here we compare actual way of usage of information with the plan of Phase 1. We may make report whether new information is useful to managements and employees or not.

Following are 10 phases of information development and show those in a diagram as Figure 3.

1) Phase 1. Strategic Information Design
2) Phase 2. Tactic Information Design
3) Phase 3. Administrative Information Design
4) Phase 4. Operational Information Design
5) Phase 5. Output Creation
6) Phase 6. Test of Output
7) Phase 7. Trial Use of Output
8) Phase 8. Use of Output
9) Phase 9. Audit on Usage of Output
10) Phase10. Survey Information Requirements

At end of Phase 1, we have review of the results and decide “go next phase” or “revise” some of document by additional works, or “discontinue” and go to “point D”. We have same procedure of review at end of Phase2 and Phase3.

If at Phase 3 review decided to go next, then it means we execute the “information design project” to come to the Phase 10.

At Phase10, information requirements are to be roughly surveyed and those are filled in the work sheet. The work sheet must include following items.

1) Requirement name
2) Timing requirement
   (Frequency / Offset / Response Time)
3) Type Information
   (Strategic / Tactic / Administrative / Operational)
4) Business purpose (to accomplish what)
5) Actions and/or business decisions
   (who will do what with information)
6) Benefits (both tangible/intangible)
7) Supporting outputs
8) Supporting process
9) Required data elements
10) 10Originator/Receiver (who and where)

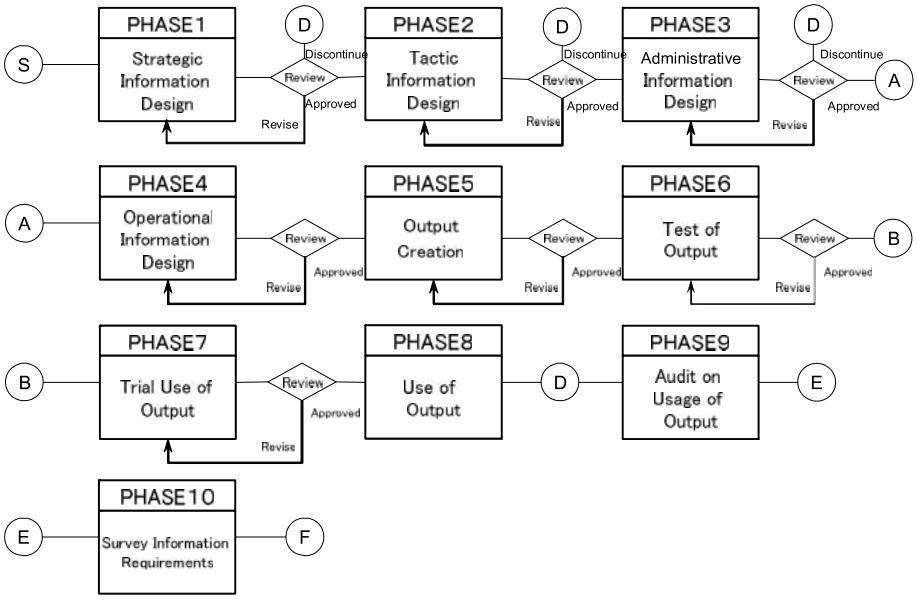


Figure 3: Information Design Methodology

11) Requirement name
12) Timing requirement
(Frequency / Offset / Response Time)
13) Type Information
(Strategic / Tactic / Administrative / Operational)
14) Business purpose (to accomplish what)
15) Actions and/or business decisions
(who will do what with information)
16) Benefits (both tangible/intangible)
17) Supporting outputs
18) Supporting process
19) Required data elements
20) 10Originator/Receiver (who and where)

Those collected work sheets are sent to Phase 1, and analyzed for study of organizational needs on information.

5 EXAMPLE ANALYSIS

After we applied the methodology at X-company, two years passed. During the time, X company collected over 2000 sheets of information requirements definition at Phase 10, but in reality they developed 11000 illustrated reports/screens at Phase 1 and at Phase 5 they developed same number of Outputs to satisfy concrete user needs. At X-company, CEO sets up his policy that information like accident/client claim is to be communicated throughout all 4 levels of organization within 10 minutes after happened. Class 1 information which is to be used by Top managements count 2200 kinds of information and still are increasing. Class 2 information which is to be used by division directors counts 1100 kinds. Class 3 information for managers counts 1800 kinds. Class 4 Information for operators and clerks and sales man counts 5000 kinds. It is interesting that they express Operational information as kitchen information. Any employee is pleased to make Output and shares information like eating curry rice in the kitchen [8].

At X-company, the evaluation of methodology is still in early stage. It might be too soon when we talk about merits and demerits of the Methodology. It is fact that X-company is recording better continuous performance in their consecutive 3 fiscal years. Structured information is easy to access by any employee and management and well maintained for users in the enterprise. Table 1 is the research result regarding needed information at each level of organization of X-company. This survey tells us there is a clear distinction between the levels. Information is structured, so it is well organized for anyone to use. The data as a digital expression of fact and event is stored in file. All employees are responsible to keep Data-Base strictly correct. The CEO believes in the enterprise there is no secret based under open management policy. Even new faces of enterprise are ready to act on adequate information. Table 2 is our survey result at X company about the degree of how much they need Information at each level of organization hierarchy. This shows absolutely that we are successful to coordinate best-fit information to the object at organizational level [9].

Finally, we can say X-company has become an information-enabled company who may be surviving in the century [10].

Table 1: Needed Information: Rank1-Rank5

Needed Rank No.1～No.5	Top Management	Division Director	Section Manager	Clerk I operator
No.1	Indus try Analysis	Technology Innovation	Market Claim	Monthly production Plan/Result
No.2	New Product	Market Share Trend	Production Quality Quantity	Weekly Work Plan
No.3	Industrial Engineering	Product Profit	Employee Productivity	Daily Quality Circle Ideas
No.4	Market Analysis	Quarterly PL/BS/Cash flow	Inventory Loss	Monthly Salary
No.5	Competitive Analysis	M&A Company List	Sales Quantity	Bonus Plan

Table 2: Degree of Information Needs at X company Questionnaire Analysis

Organization / Characteristics	Top Management	Division Director	Section Manager	Clerk or Operator
Degree of Dependency on External Source	Very High	High	Low	Very Low
Degree of Dependency on Internal Source	Medium	High	Very High	Very Very High
Long Ranged Trend Analysis/Estimated	Very High	High	Low	Very Low
Narrative Historical Analysis	Low	High	Very High	Very Very High
On-Line Real Time Information	High	High	Very High	Very High
Current Performance Analysis Information	Medium	High	Very High	Very Very High
Periodic/Cyclic Reported Information	Medium	High	Very High	Very Very High
Simulated Analysis Information	Very High	High	Low	Very Low

6 CONCLUSION

Soon after the real apply of the methodology at X-company, we suggested to use the methodology at multiple enterprises. Those enterprises reported about good points of real use of the methodology. Following are some comments from those enterprises.

- Bird-view visibility has increased from top management level to operational level.
- Information used at upper level are very reliable since the basic information like operational/administrative information are reliable.
- Information requests from upper level managers and directors and managements are severely examined through the Phase 1, 2, 3. operational information does not randomly increased.
- Literacy of information utilization is increasing at enterprise.
- It is still too early to conclude the methodology has the absolute merits.

REFERENCES

[1] Managerial information, Bearingpoint Co., Ltd. Seisansei Press (2005).

[2] Paul A. Strassmann, Information Payoff, The Free Press (1985).

[3] Milt Bryce and Tim Bryce, The IRM Revolution: Blue Print for the 21st Century , MBA Press (1988)

[4] Kazuya Matsudaira, et al, New Methodology for Information Resource Management by 4 aspects Parallel Approach, in Proceeding of KES2007 (2007).

[5] Hope Jeremy and Robin Fraser, Beyond Budgeting-How managers can break free from the annual performance trap, Harvard Business School Press (2003).

[6] Peter F Ducker, Management Tasks Responsibilities Practices (1974).

[7] R. S. Kaplan and D. P. Norton Strategy Focused Organization: How Balanced Scorecard Companies Thrive in the new business environment, Harvard Business School Press (2001).

[8] Rober Simons, Performance Measurement & Control Systems for implementing Strategy, MacGraw-Hill (2003).

[9] Howard E. Aldrich, Organizations Evolving, Sage Publications Limited (1999).

[10] Ben Gilad, Early warning - using competitive intelligence to anticipate market shifts-, control risk, and create powerful strategies, AMACOM (2004).

Remote consultation system using hierarchically structured agents

Hiroshi Yajima*, Jun Sawamoto**, Kazuo Matsuda***

* Faculty of School of Science and Technology for Future Life, Tokyo Denki University, Japan
**Faculty of Iwate Prefectural University, Japan
*** School of Science and Technology for Future Life, Tokyo Denki University, Japan
yajima@im.dendai.ac.jp, sawamoto@iwate-pu.ac.jp

Abstract -In fields of technological innovation the speed of advance is fast, and while it is difficult for some people to keep up, there are few experts in new technologies. Since consultation is focused on a small number of experts, phenomena such as being unable to obtain sufficient information in a timely manner occur, and are one of the major reasons for the increasing social-technological divide. This paper proposes a 2-level hierarchical remote consultation system using two types of agent. The system possesses the features that through the responses to consultation made in advance by multiple agents, experts can focus on only complex questions, and in addition, consultees' waiting times are reduced. Its effectiveness is demonstrated experimentally.

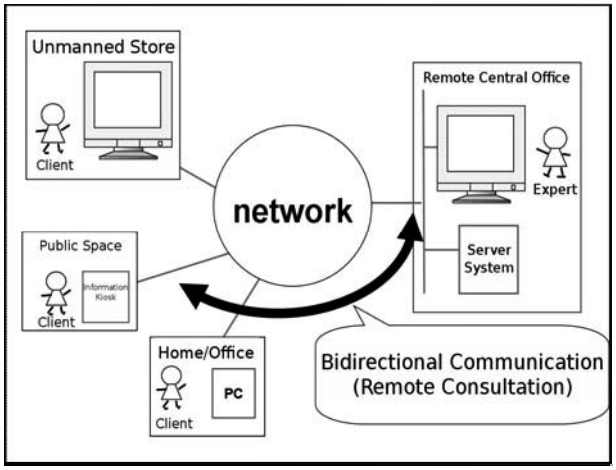
Keywords: remote consultation system, agent, remote communication, expert, TV conferencing.

1 INTRODUCTION

Society has been aging in recent years, and service functions for poorly informed aged persons and patients will be sought. While the number of healthcare professionals is small, remote healthcare consultation which is efficient and yet maintains an appropriate level of service is being sought. In addition, while forms of employment are diversifying, models of employment such as the teleworking remote office are gathering attention. Further, in fields of technological innovation the speed of advance is fast, and while it is difficult for some people to keep up, there are few experts in new technologies. Since consultation is focused on a small number of experts, phenomena such as being unable to obtain sufficient information in a timely manner occur, and are one of the major reasons for the increasing social-technological divide according to which the benefits of advancing technology cannot be fully realized. Regarding policies for resolving this issue in society at present, research focusing on the theme of efficient remote communication support is important. In particular, support for fostering communication among disparate groups of fellow persons is essential.

So far, remote consultation systems have been conducted via TV conferencing and so on [1,2,3,4]. However, in such cases, it has been usual for consultation to be conducted with 1 consultee exclusively occupying the services of 1 expert. Regarding information sharing there are also remote conferencing systems [5] such as Skype. Functions for visualizing the topic of a discussion among its members in a shared manner have also been proposed. However, remote conferencing has been centered on discussions along a common theme among all members, and they are inefficient for situations in which experts possessing knowledge and information in a given field present solutions to laypersons lacking such information.

In order to solve these problems, this paper proposes a 2-level hierarchical consultation model using 2 levels of agent. The two types of agent established are Service Agent systems (SA) for the clients, and Supervisor Agent system (SVA) existing between the SAs and experts, who provide easily understood support by responding to requests for support from SAs in cases when they are able to do so, or otherwise forwarding the existing message history to experts. This system possesses the features that experts are able to focus on only complex questions, and in addition, consultees' waiting times are reduced.



2 PROBLEMATIC POINTS

Figure 1 Existing model of a remote consultation system

2.1 Existing Remote Consultation

Remote consultation operations over the internet are increasingly tending towards communication among people from different cultures and institutions. This is because the internet generation, new technologies, new organizations and new establishments are being developed, constructed and disseminated on a daily basis, and it has become necessary to rapidly assimilate this flow.

In remote consultation, there are synchronous and asynchronous models. Synchronous models are those such as a telephone, where both parties exchange discourse during the same period of time. Asynchronous models are those such as email in which discourse may be exchanged without adopting a specific time period. Asynchronous models are mainly being applied by means of email, but with the rapid speed of business in the present day, there is an increasing need for synchronous models. The objective of this research is a synchronous remote consultation system over the internet among these kinds of disparate groups and individuals.

Figure 1 shows an example of an existing remote consultation system which has already been investigated [6,7]. Basically, consultees initiate consultation from a convenient location, while on the other hand, a small number of experts oversee these consultations from a central office and respond to complex queries.

Remote consultation is currently being conducted in many fields. PC user support and so on, is widely active in general. Also, remote consultation has also come to be provided in financial and healthcare fields.

Along with this model, the provider model has also diversified. At present, remote consultation services are being provided by email, homepages, TV, telephone, and models combining these technologies. However, services

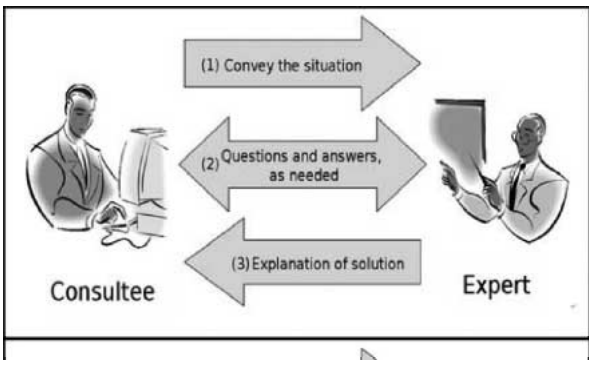


Figure 2 Phases of remote consultation

using asynchronous communication models such as email incur a time-lag between the receipt of a consultation and the response, so problems cannot be solved immediately. For reasons such as this, the telephone, with its synchronous communication model, is the main channel for the provider model of remote consultation services.

It is thought that the general flow of consultation may be broadly divided into 3 phases [6], and in this research the following definition is adopted (see Figure 2).

2.2 Problems With Existing Systems

Remote consultation has the following features (communication patterns). Basically, partners from different cultures (clients and experts) communicate as fellows. The disparate groups may include for example, a) groups of experts and laypersons, b) intradepartmental and interdepartmental staff groups, and c) groups of company staff and non-company persons, which thus constitute groups of people with different values, knowledge and objectives. People belonging to heterogeneous cultures often have different levels of knowledge, and the range and content of their basic assumptions also often differ, yielding obstacles to communication.

Also, the number of consultees is usually overwhelmingly greater than the number of experts, so if experts respond to consultees on a 1-to-1 basis, the efficiency of consultation is poor.

3 SOLUTION STRATEGY

3.1 Concept

This paper proposes a formula for conducting remote consultation in which experts and agent systems are combined. Consultation is therefore first conducted between consultees and service agent systems (SA), and the SAs are supported by experts in the basic model proposed. This allows consultation to be conducted between consultees and agent systems, without the need for 1 to 1 consultation between consultees and experts.

Next, the multiple SAs seek support from the experts in cases when they are unable to respond themselves. However, when multiple SAs seek support simultaneously, experts must deal with multiple support requests at once. Agent system (SVA) with different functions (meta-knowledge and scheduling functions) is therefore placed between the experts and the SAs. By constructing the agent system in 2 layers (SVAs and SAs), consultation is made efficient. By including the SVAs, experts need only deal with a single SVA, rather than multiple SAs.

3.2 System Structure

The consultation model of this research is shown in Figure 3. Adopting this structure gives rise to the following advantages.

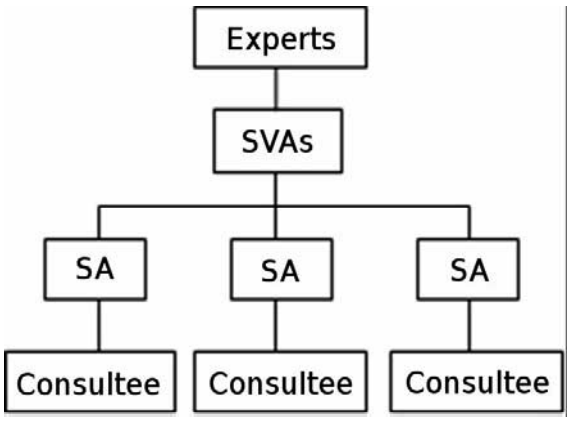


Figure 3 The model proposed in this research

- The problem arising when multiple SAs directly request support from experts simultaneously, thus increasing the burden on experts and decreasing the efficiency of consultation, is avoided.
- Also, the problem associated with consultee stress arising when multiple SAs send requests for support simultaneously, and one SA must wait for another SA's support to be concluded, thus increasing their consultee's waiting time, can be solved.

The function of each agent is as follows.

SA: conducting information exchanges with consultees. In this research, SAs question consultees regarding essential items and obtain their replies. When SAs are unable to respond themselves, these replies are forwarded to SVAs as requests for support.

SVA; providing support for experts, acting between the SAs and experts. When requests for support from SAs are within the range they can respond to, SVAs respond themselves, and in cases when they cannot respond, the requests are scheduled according to importance, and presented to experts in an easily understood manner along with the message history to date.

4. Remote Consultation Utilizing Hierarchical Agents

4.1 Processing

The following procedure is proposed as a method for realizing the concept.

(1) Consultation is promoted between SAs and consultees. SAs ask questions of the consultees, and the consultees return their replies. Only the SAs respond during this process, without involving the experts.

When the consultees' replies are correct, the SAs present the next question.

(2) The SAs send the consultees' replies to the SVAs, and the SVAs process the data, presenting individual SA consultation cases to the experts. Under this process, the experts only observe the data reported to them.

(3) When replies from a consultee incur exceptional handling, SAs request support from SVAs, i.e., when the content of replies from consultees cannot be processed by SAs, SAs request support from SVAs.

(4) When requests for support received by SVAs can be handled using the meta-knowledge they maintain, they return replies to the SAs. When they are unable to reply themselves, the preceding message history is attached, and support is requested from the experts.

(5) Experts receiving requests for support send replies to the SVAs.

(6) Messages from the experts are sent, via the SVAs, to the SAs originating the requests, and presented to consultees. After receiving these messages, SAs resume questioning.

(7) When consultees are satisfied, consultations are concluded.

4.2 Specific Flow Of Consultation

Existing consultation systems have mainly advanced using audio, but in this research, audio is not used. Consultation is conducted using a chat format in free text. The consultation advances as the agent poses questions to the consultee, and the consultee returns the answers, or asks questions. Then, when the consultee is satisfied, the final result is displayed at the consultee side and the consultation ends. The flow of consultation in shown in Figure 4.

5 EXPERIMENTAL ASSESSMENT

5.1 Experimental Objectives

In the remote consultation system using agents, the case when SAs and SVAs are utilized, and the case when only SAs are utilized are compared and the variation in the burden on experts is ascertained. The number of consultees for each SA is taken to be 3.

5.2 Experimental Conditions

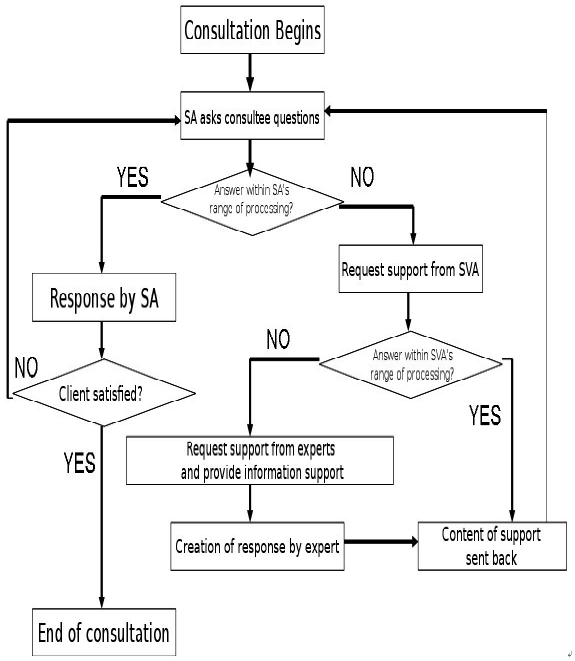


Figure 4 Flow of consultation

Condition 1 Consultation is conducted with an agent system in which 3 SAs respond to the 3 consultees. Experts respond to all of the requests for support from the SAs.

Condition 2 Consultation is conducted in a hierarchically structured system with an SVA added for the 3 SAs. SVA automatically reply when they are able to respond using their own knowledge, and send the problems to which they cannot respond, as requests for support along with the preceding message history, to experts.

Consultation is conducted using only text, without audio, in both Conditions 1 and 2.

5.3 Experimental Task

As a task, consultation was conducted regarding the bureaucratic procedures involved in registering for a new insurance policy. Consultees do not have any knowledge, and ASs have procedural knowledge, and SVA has exceptional knowledge.

5.4 Experimental Subjects

As experimental subjects, there were 1 expert and 5 groups of 3 consultees, making a total of 16 people. The subjects were students, and all had experience using a PC.

5.5 Experimental Results

(1) Data

In the experiments, as an indicator for measuring the burden on the expert, the expert's operating time was determined. The expert's operating time is shown in Figure 4. The average operating time of the expert in Condition 1 was 1570 seconds, and in Condition 2, it was 1150 seconds, so when SVAs were included, the result was a drop of about 27%.

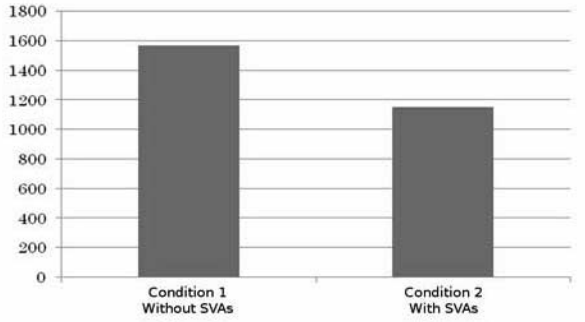


Figure 4 The average operating time of the expert in

Also, the total number of messages to the expert was 245 in Condition 1, and 117 in Condition 2, so the result was a decrease of about 47%.

Regarding the consultees' waiting times, these were measured as the period during which they could not conduct their own operations, i.e., the processing time of each consultee's agent, and the expert's operating time. The experimental results are shown in

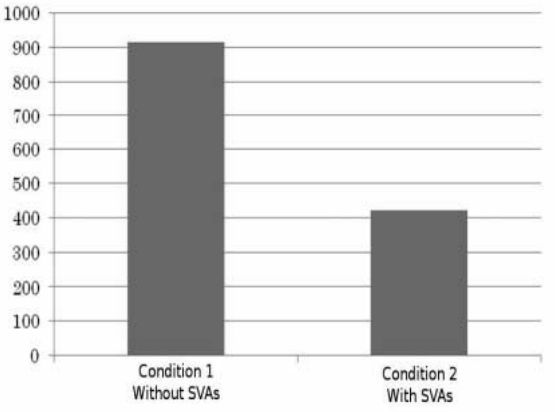


Figure 5 Consultees' average waiting times in each condition

Figure 5. The waiting time in Condition 1 is 915 seconds, and in Condition 2 it is 412 seconds. This result is a drop of about 57%.

(2) Questionnaire

A questionnaire was completed after the experiment, by both consultees and experts. The experiment was evaluated on a scale of 1 to 5 (1 was best, and 5 was worst). The results of the questionnaires are shown in Tables 1 and 2.

Table 1 Results of the consultees' questionnaire

	Without SVAs	With SVAs
Smoothness of consultation	3.3	1.6
Level of concentration	3.3	2.6
Atmosphere	3	3
Reliability	2.3	2.6
Ease of consultation	2.6	2.3
Degree of stress	3.3	2.6
Level of satisfaction	3	2

6. DISCUSSION

6.1 Expert's Operating Time

Looking at the expert's operating time, when SVAs are present the time is reduced in comparison to when SVAs are not present. It was thus proven that the presence of SVAs reduces the expert's burden.

However, there is a big difference in the reduction of the expert's messages by 47% as compared to the reduction of 27% in operating time. The content of the questions directed at the experts is therefore classified in Table 3.

Table 2 Results of the experts' questionnaire

	With SVAs	Without SVAs
Ease of use	3	3
Ease of information acquisition	3	2
Level of concentration	2	3
Level of stress	2	3

According to this data, it can be seen that when SVAs are present, there is a reduction in questions regarding phrasing whichpresent, there is a reduction in questions regarding

Table 3 Total number of questions in each classification

Evaluation Items	With SVAs	Without SVAs	Average response time (seconds)
1. Phrasing	77	62	17.055556
2. Price	30	32	31.875
3. Service	6	15	33.125
4. Personal circumstances	16	18	32

phrasing which do not require the experts long to answer, and an increase in other types of time-consuming question, particularly those regarding the service. According to the consultees' post-experiment questionnaire, consultation is smooth when SVAs are present, which means that there is an environment in which it is easy to ask questions. It was thus understood that while there are individual differences, making the consultation smooth may increase the consultees' motivation to ask questions.

6.2 Consultees' Waiting Times

Looking at the results regarding the consultees' waiting times, the waiting times are reduced when SVAs are present, in comparison to the case when they are not. It was thus understood that consultees' waiting times may be reduced through the use of SVAs.

6.3 Questionnaire results

Looking at the results of the questionnaire, as shown in Table 1, the consultees' overall evaluation is increased when SVAs are present. In particular, the evaluation of the smoothness of consultation is very much increased. However, while the overall evaluation is increased, the evaluation of reliability is decreased. According to the post-experiment questionnaire, this means that there is a little resistance to the fact that the responses to questions come from a computer. It was thus understood that in contrast to the increase in the efficiency of consultation, there is a demerit in the sense that the reliability ends up decreasing.

Looking at the results of the questionnaire shown in Table 2, when SVAs are present the expert's ease of acquiring information is increased, so it can be seen that consultation has also been made easier for the expert. However, the evaluations of the degree of stress and level

of concentration are decreased. This is thought to be related to the fact that the expert's operating time is decreased, so their free time is increased, which may affect their levels of stress and ability to concentrate.

7. CONCLUSION

This paper proposed a 2-level hierarchical remote consultation system with 2 levels of agent. The proposed system is established with SA agents who respond to clients, and SVA agents existing between the SAs and experts, who respond to requests for support from SAs when they are able, or if not, request support from an expert by sending an easily understood request along with the preceding message history. Experimental evaluations proved that the establishment of SVAs shortens experts' operating times, and that the system is applicable as a one-to-many remote consultation system.

REFERENCES

[1] S.Matsuura, T.Matsumoto, Y.Kiyosue, T.Sugawara and S.Masaki, "Development and Evaluation of the NetForum Simplified Multipoint Television Conferencing System", Proceedings of the Information Processing Society of Japan, Vol.41, No.11, pp.3142-3151 (2000)

[2] S.Kobayashi and T.Iwaki, "Home Teleworking Experiment using a Multipoint Internet Conferencing System", Technical Report of the Institute of Electronics, Information and Communication Engineers of Japan, OIS2002-11 (2002-5)

[3] T.Obata and R.Sasaki, "OfficeWalker: Video Image Transmission System Supporting Incidental Conversations in Distributed Offices", Proceedings of the Information Processing Society of Japan, Vol.40, No.2, pp.642-651 (1999)

[4] S.Ido and K.Inoue, "Development of a Video Conference Minuting System Based on Nodding", Technical Report of the Information Processing Society of Japan, Vol.34, pp.19-24 (2006-3)

[5] D. M. Grayson and A. F. Monk, "Are you Looking at Me? Eye Contact and Desktop Video Conferencing", ACM Transactions on Computer-Human Interaction, Vol.10, No.3, September 2003, pp.221-243.

[6] T.Tanaka, Y.Koizumi and H.Yajima, "Proposal of a Dialogue Support Environment for Remote Consultation using an Unmoderated Communication Mode", Proceedings of the Human Interface Society, Vol.4, No.3, 2002

[7] T.Tanaka, H.Mizuno, H.Tsuji, H.Kojima and H.Yajima, "Remote Consultation System Supporting Asynchronous Communication in Distributed Environments", Proceedings of the Information Processing Society of Japan, Vol.40, No.2

A Proposal for Business Process Models for Mashups Especially for Enterprise Systems

Nobuhiro Kataoka
School of Information and Network, Tokai University
E-mail: kataoka9@keyaki.cc.u-tokai.ac.jp

Abstract-Mashups are now widely used as a method of freely offering new services by combining various services that already exist on websites. However, they have not sufficiently penetrated enterprise-level utilization yet. This paper first outlines the current mashup situation. It then classifies how mashups are used in enterprises. Further, it analyses the features of the business process model for each form of usage. Based on this analysis, it makes suggestions on the usage of mashups in enterprises. Finally, the author refers to issues regarding the promotion of mashups in enterprises.

Keywords: mashups, business process model, WebAPI, enterprises

1. INTRODUCTION

Tim O'Reilly listed seven items as the basic principles of Web 2.0 in his paper "What Is Web 2.0" [1], among that, we consider, mashups will play a great role in the following items:

i) Highly cost-effective and upgradable services to be offered, rather than packaged software which is not always user-friendly.

ii) Utilization of collective intelligence

iii) Adoption of lightweight user interfaces, lightweight development models and light weight business models.

Mashups are a way of using collective intelligence, and developments that use mashups are so high in development efficiency and can create lighter development models.

As a typical example of a Web API, it is possible to enjoy different Google services by means of Google API (we call it a Web API to discriminate it from Java API which uses functions in its program, whereas Google uses functions on the network).

There are a number of Web APIs now in service. Utilizing these services in a better way will help us develop our targets in much shorter time and at much lower cost than conventional development. We refer to services that use two or more Web APIs with the addition of some programming as mashup software, and we can see a number of them on the Internet. In a sense, it is a kind of system development using the assembly of services. With conventional techniques, we developed systems which there should be to create. On the other hand, mashups start a survey of the intended development: what services (Web APIs) can we find? This means that the world has changed itself so that a new software can be assembled with a set of services. In a sense, it is a prominent breakthrough, and the paper discusses how the world of enterprises should take in such services.

2. Present situation of mashups

A mashup is an application which develops a system by combining Web APIs. Mashups use the following two methods for development:

-REST: In this method the information that is sent to the service offer side: URL+ demand item of one's home page

For example, searching of an article Kaku.com is next form.

ttp://api.kakaku.com/Ver1/ItemSearch.asp?Parametername1=value, parameter name2=value,,,

The return data are in XML format

- Web service: This method use SOAP.

And this is same usually Web service.

Most Web API support the REST method, and the most convenient point of REST is that we can make mashups with only a home page. There is no need to create a server program.

In May 6, 2008, there were approximately 3000 mashup sites, which was an increase of 450 during the previous six months [2]. In services, mapping accounts for 33% of mashups and photo, video, and music account for 10% or so each.

Google Maps accounts for about half of Web APIs followed by Flickr, Amazon and YouTube with a ratio of 10% or so.

In addition, tools to help mashups which click on Web APIs with a mouse on the screen are being developed for example [5] and thus the development of mashups is increasing more and more.

3. Effective use of mashups in enterprises

Most of the current examples of mashups are, in general, related to searches of information for maps, accommodation, food, traffic routes, images and others, that is, closely related to daily life or people's hobbies. In other words, at the moment utilization by enterprises lags behind the examples given above. The term "enterprise"

Table 1 Popular Web API for databases

Vendors	Site name
Amazon	Book information etc.
East	Route search etc.
Kakaku.com	Merchandize offered by Kakaku.com
Recruit	Information and catalogs of used cars
JAL	Booking service for corporations
JTB	Travel services that JTB offers

used here refers to corporations, associations, government offices and local public bodies.

When an enterprise is going to utilize mashups, they should consider from two viewpoints: one is that the enterprise stands on the side of providing Web APIs, and the other is that the enterprise is on the side of users of Web API. A significant and special characteristic for both viewpoints are that a databases are involved in some relationship in both case.

3.1 Offer of Web API by enterprises

As Amazon is now performing, the purpose of this offering is a strategy to induce other companies to their side while offering a Web API to allow access to its own database.

Today a number of corporations use APIs, some of which are shown in Table 1.

In the current situation, most mashup users are in the service industry, where manufacturing companies provide Web APIs that relate to their product database, and parts supply companies provide Web APIs that relate to the database of their components, and thus it is considered that the sales companies that deal with such database are enabled to build effective business models.
Besides such products databases, we can assume information control Web API performed by some other third parties.

A specific example is given below.

(1) Traceability center

Today, the traceability of the place of production of foodstuffs, their processors, additives and other related matters is demanded for food safety. This is not limited to groceries, but applies to industrial products that are manufactured by combining many parts. For these requirements, this is a form to establish a traceability center of third parties to control the IDs of raw materials and parts used, and each processor and manufacturer to enter trace information into the database of that center via a Web API.

(2) Customers information control center

As a result of the enforcement of the Personal Information Protection and Electronic Documents Act, companies are paying a substantial amount for the control of personal information. Therefore, it would be better for the company to leave the task to a trustworthy professional company. When it is required to use personal information, this form will allow the user company to retrieve necessary data by means of a Web API.

3.2 Use of Web API in enterprises

If we categorize the usage in enterprises by two axes: an axis of whether it is the utilization of external services or in-house services; and another axis of use by customer or inside of the company, then we get the results shown in Fig. 1.

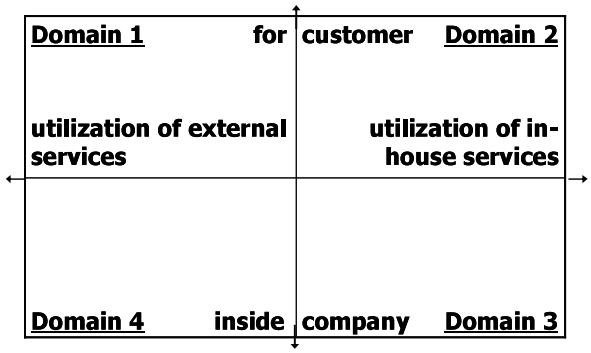


Fig. 1 Classification of using mashups

(1) Domain 1

Mashups in this domain offer external services to customers using Web APIs. In this case, external services as well as internal services in the enterprise undergo mashing.

If a company used information from Kakaku.com, and added the company's own services and offered the results to customers, for example, that would fall into this category. Most cases that use Web APIs listed in Table 1 belong to this domain.

The portal site Rearden commerce service[13] for business tripper procurement possesses an external service for hotel booking, air ticket booking and an internal service for travel expenses settlement, and it has been unfolding as SaaS, but if we see it as a system, it fall into this domain [4].

If interfaces are different between electronic application and a civilian service, mashup could unify these interface and offer convenience for user in some cases [10].

(2) Domain 2

What is positioned in this domain is offering internal services to the external world, and it is considered to correspond to the offer of Web API mentioned in 3.1.
More specifically, some businesses provide customers with a bundle of services of Customer Relationship Management (CRM) and Customer Satisfaction Management (CSM) and other support systems via a

portal site [11]. Depending on the cases, this may function as a portal site for a customer service staffer of corporations. In such cases, it belongs to domain 3.

One other alternative is what offers services for mission critical systems, and in this case, it can exist only for a provider of Web API.

(3) Domain 3

This domain includes mashups of two or more in-house services that have been united into one by a mashup in order to offer the results to in-house users. This is represented by an example as follows:

The operator support system of the KDDI call center consists of two or more internal and external databases, which have been mashed into one. A Kapow mashup Server developed by Kapow Technologies, US is used [12].

DeeCorp and Relocation Japan reflects the customers data to there own system by using Web Service APIs ,which is controlled by SAP Business One. Salesforce.com services SAP system by SaaS. Furthermore at the next stage they are planning to reinforce the mashup with marketing information offered by Google Maps and Yahoo [4].

(4) Domain 4

This domain offers services for employees, where "Shuccho JAWS" (comically meaning, "Good business trip!" in Japanese) [3] can be said a typical one of them. This service works in liaison with an internal service, called Travel Expenses Settlement, along with various types of external services.

Together with similar ones, LMS (Learning Management System by network) is based on the system of a corporation that specializes in professional education for many company's employees, and it is also possible to take the results into the data of the company's own education systems.

A certain university created a mashup of the university own mail form on the basis of Gmail without building a school email system independently, which is also a typical case.

It may be supposed that corporations will be extremely resistant to leaving their electronic email to control by other parties, but Google has already started a paid Gmail (Google Apps) service in which the security functions are enforced, and therefore, it is anticipated that an increasing number of corporations will relinquish control of their email servers in the future. Here various Web APIs are offered, and it will be easy to add the corporation's unique functions.

4. Proposal of basic concept for development of a mashup

The greatest advantage of mashups is the point that is joining of services only by making of JavaScript. So this is only possible by make home pages and no needs to make server programs. But it is necessary to take a method corresponding to each use case.

(1) Case of no close relationship between external and internal services

To avoid cross-domain limit between external services used, use an "src" option of <script> tag. This is similar to the normal creation form of a mashups.

When putting external services and internal services into cooperation, prepare Web API interface for the internal services, and mashups including this Web API.

(2) Case of requiring close data coupling

If close coupling is required between an external service and an internal service, it is needed to make internal service that receives two or more external services.

(3) Case of using an external service to work as a database

It is necessary to use the services of a trustworthy party. This is a case of receiving a service that offers information from merchandize databases, selling air tickets and other data. In the current situation, such services are conducted by reliable major corporations, and it is not likely to lead to a problem.

(4) Case of using system services for mission critical systems of a company via an external service.

In such cases, it is necessary to sign a Service Level Agreement contract (SLA). Depending on the case, a function may be generated that stores the data in the company's database.

5. Details of development of a system based on the proposal

5.1 Features of business process of using mashups

- The use of mashups enables us to easily take in external functions. In other words, we can simply build a business process that is complicated as a result. However, the mashed up business process of the other party will remain a black box.
- We can easily and widely modify the service of the system by changing the party for mashup.
- In the mashup of an external service, we do not have to consider the SLA so strictly if the subject is a support system (i.e. there will be no critical influence on corporation activities even if the information system, map, route guide or other systems are suspended).
- It is necessary to exactly evaluate the SLA if the external service for the mashup deals with a merchandize database or the like, but such services are often offered by trustworthy corporations in the current situation, and this will not likely lead to a problem.
- If the external service for a mashup is a mission critical system, it is necessary to evaluate the SLA in detail.

- A service can be considered available that mashes tasks of mission critical systems of a corporation and that offers the results to the corporation under an SLA that is suitable for users.
- Ad-hoc system is a different variation that meets the characteristics of the corporation, and this is considered outside the scope of mashups. However, it can be possible to mashup small and analytical parts, instead of services.

5.2 Clarification of business process of proposal form

The following section uses BPMN to describe business process models. The BPMN description will be conducted by domains of the classification of Fig. 1, focusing only on how information should be exchanged between lanes or between pools although detailed expression is omitted.

However, what is used by BPMN is only the concept of lanes and pools. A pool expresses an external service, and a lane expresses the relationship between internal services.

(1) Domain 1, No. 1

This case masheup offers external services for support systems and internal corporation services. Since the external service is for support systems, its temporary suspension will not always be fatal to users. Those services connect home page functions to each other using JavaScript.

It is also possible to connect internal services to each other using mashup by setting the Web API functions in a similar way to the external functions.

Since a Web API that has similar functions is very likely to exist, it is possible for the service of support systems to take building-in of a switching function into consideration.

However, check the time setting to ensure no response is found, and if you are going to set an alternative to any other service, it is needed to make connection with internal components that performs such checking.

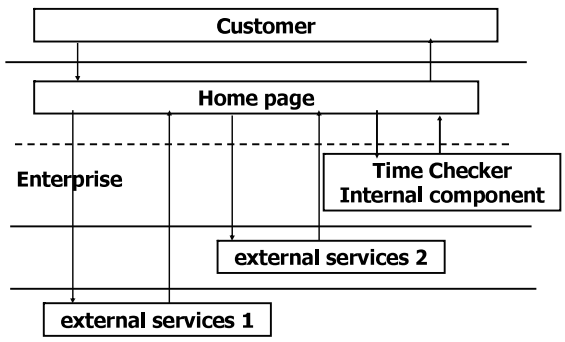


Fig.2 Domain 1-1 BPMN presentation

(2) Domain 1, No. 2

This is to offer mainly external services for mission critical systems to general customers, with some internal services added. In this case, most services are information related to a database as an external service. Utilizing that information of database, and adding internal services, thus services are offered to customers.

Such internal components will become necessary to perform the control of customer information after securing external services of merchandize information, for example.

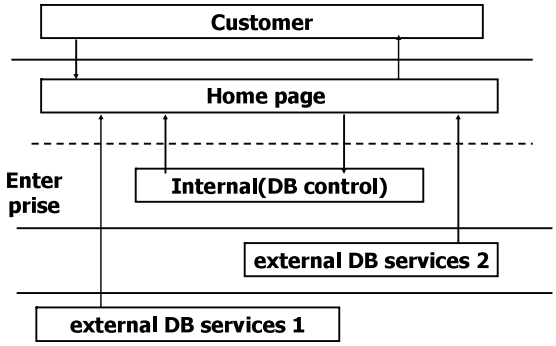


Fig. 3 Domain 1-2 BPMN presentation

In other words, this form not only merges what was taken from two or more external services using JavaScript on the homepage, but also offers them to users after processing by internal services.

(3) Domain 2, No. 1

This is a portal site for in-house users, and basically this does not exist as Domain 2 for customers basically. But it exceptionally exists only when it offered directly to customer's in-house information.

This builds a portal site for user support divisions through a mashup of systems (services) found throughout the company. This is also a method for customers themselves to mashup through releasing the Web API. For examples of mashup tools of in-house systems, see [8] and [11].

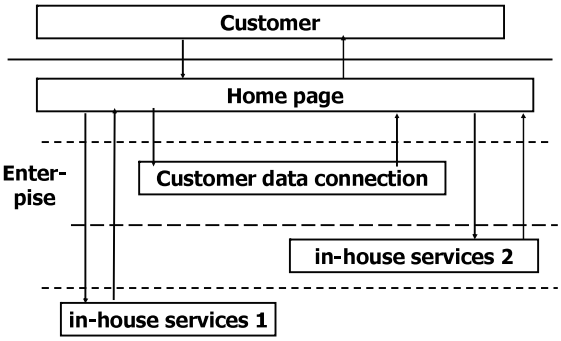


Fig. 4 Domain 2-1 BPMN presentation

This is for customers to check their registration information by themselves.

(4) Domain 2, No. 2

This is offer of a service that takes various in-house data as its nuclei. We can consider it as corresponding to Table 1. This is an offer of Web APIs to external users; it does not belong to a mashup domain.

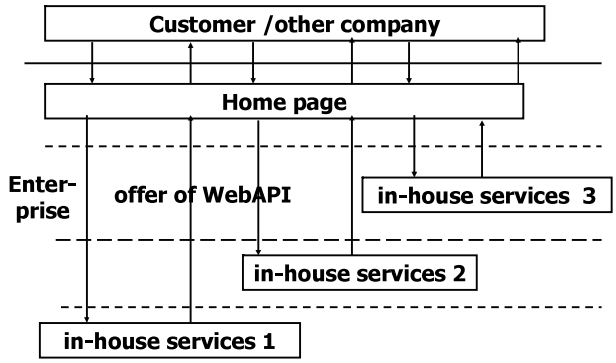


Fig. 5 Domain 2-2 BPMN presentations

(5) Domain 3, No. 1

This includes cases that integrate internal services and works as a portal site to provide in-house users with customer information which dispersed in the company. The tasks in this case are often support tasks such as customer service (call center). This requires an internal service to integrate information.

Other relevant tasks would include tasks of Customer Relationship Management (CRM) and Customer Satisfaction Management (CSM).

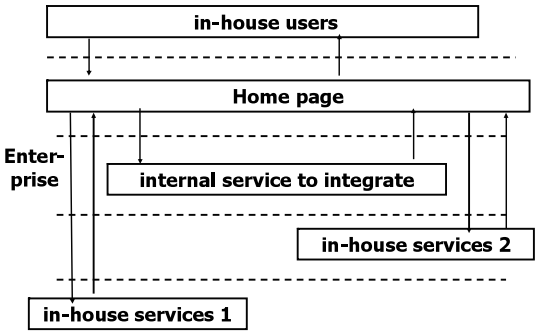


Fig. 6 Domain 3-1 BPMN presentation

(6) Domain 3, No. 2

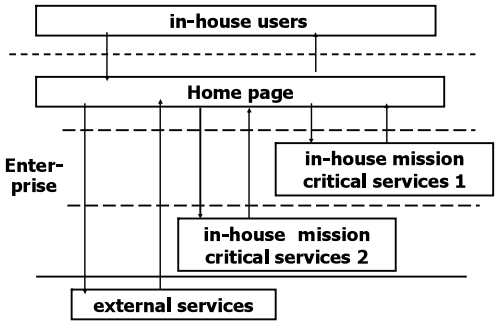


Fig. 7 Domain 3-2 BPMN presentation

This mashupes mission critical system found within a company. Consistency of a database is demanded in this case. Therefore, it is not suitable for Mashup.

(7) Domain 4, No. 1

In this case, a corporation couples external services and internal services together and offers the product to in-house users. "Shuccho JAWS" is a typical example.

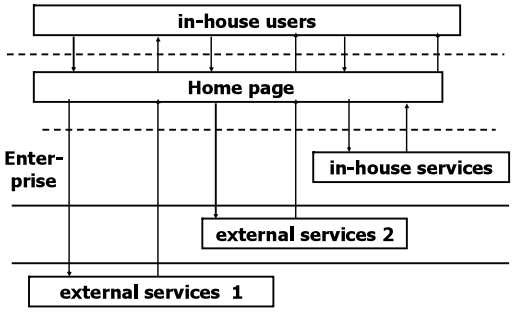


Fig. 8 Domain 4-1 BPMN presentation

In-house services include a system of settling business trip travel expenses. The external services in this case, however, are supporting services such as route guides and accommodation guides. If there is no internal service, it eventuates in a normal mashup; it has no relation with any enterprise systems.

(8) Domain 4, No. 2

In this case, a corporation mash ups external services and if necessary, internal services as well, to offer the results to in-house users. The difference from domain 4 No.1 is that this external service deals with basic key tasks of the corporation. This should include a mashup of an ERP package, for example. External services such as a partially altered SaaS user interface can be one of them as well.

Here a case can be possible of performing control functions of basic key database of external services as an internal service, as in domain 1 No.2. In a system that involves heavy traffic, however, the use of such an external service is required as SLA can be concluded on it.

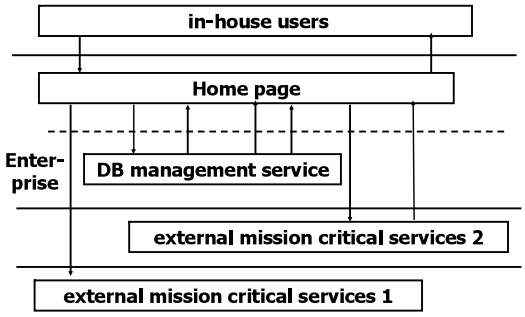


Fig. 9 Domain 4-2 BPMN presentation

6. Process of development in mashup

According to the business process in Section 5, the steps given below enhance the development efficiency by means of the practical use of mashups.

(1) Clearly define the reason for development.

This is never different from conventional process of development; it is the most important matter in development. Creation in an enterprise produces a big difference from general mashup creation that is a result of a sudden idea.

(2) Decision of functions demanded and decision of which mashup to use

When the required functions have been decided, which kinds of Web API exist and which are available should be investigated. When doing this, some community sites such as Programmable Web [2] are effective as mentioned previously. Further, use of search, comparison and matching sites of Web API is also effective [9].

(3) Creation of skeleton diagram of business process

The whole process of business should be clarified. There is no need to create a detailed business process diagram here, and it would be sufficient to specify only which Web API to use and what information should be exchanged, and at this time point which of domains given in Section 4 the subject for development belongs to should be recognized, and securely take in the conditions required there.

(4) Developments in the mashup environment

Currently two or more development environments have already been provided by vendors, and the utilization of these environments has been discussed [6-8].

In addition, the systems developed for in-house use take the form of offering a Web API, utilizing the interface, and developing relevant systems.

7. Measures to promote mashups

To promote mashups, various themes are found, but I list only the themes here, of which details will be released in some paper in the future.

(1) Software companies aim at becoming a business that opens their packaged software by means of a Web API; they aim to sell services, rather than packaged software.

(2) The elements for promotion of mashups are to clarify the authority to use Web APIs and mashups and to establish the form of clarifying mutual relationship between them.

(3) Service level agreement (SLA)

First, it is necessary for this agreement to be positively concluded, but it is necessary first to determine whether the other party is suitable for the tie-up.

On the other hand, another big issue is to find how many Web APIs exist, and there quality level. To solve this problem, some sites that collect and release users' evaluation will play an important role.

8. CONCLUSION

Enterprises have already started to utilize mashups. In this situation, we have tried to sort out current factors in the light of process of development system and business process models where mashups are used. Our intention has been to clarify them in order to identify what are the development system issues. In this comparison, we used two axes for classification: an axis of whether the user is inside the organization or is a customer, and the other axis of whether the service used (database) is outside or inside the organization; and discussion was conducted by four quadrants produced by the two axes.

As a result, the items to be clarified at the development stage have been identified for each domain.

REFERENCES

[1] Tim O'Reilly, What Is Web 2.0, http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[2] http://programmableweb.com/

[3] http://bic.recruit.co.jp/index.html

[4] Nikkei Computer, 2007/5/14, pp. 43

[5] Rob Ennals, Minos Garofalakis MashMaker: mashups for the Masses, SIGMOD '07, June 12-14, 2007, pp. 1116-1118

[6] Google Web Toolkit http://code.google.com/webtoolkit/

[7] WebSphere Portal for Google Gadget http://www-01.ibm.com/software/brandcatalog/portal/portal/details?NavCode=1WP1001AJ

[8] Oracle WebCenter Suite http://www.ebizq.net/news/7386.html

[9] http://www.api-match.com/

[10] Fukui Hiroki et al. Alternative Way for Government Private Portal Using mashup Technique; IPSJ SIG Notes, Vol. 2007, No. 16 (20070301) pp. 19-24

[11] Data mashups online service http://www.datamashups.com/

[12] Kapow Technologies http://www.kapowtech.com/products/products.aspx

[13] Rearden commerce http://www.reardencommerce.com

Session 3: Education System (Chair Kazuya MATSUDAIRA)

Proposal of program learning tool using execution histories

Eiji Sugino*, Ryusei Uchimura**, Jun Sawamoto* and Norihisa Segawa*

* Faculty of Software and information Science, Iwate Prefectural University, Japan
**Infocom Corporation, Japan
{sugino, sawamoto}@iwate-pu.ac.jp

Abstract - Nowadays, we can easily obtain a copy of practically any used program in our open source community for learning. However, the reality is that the level of such practically used programs is often complex and of such a large scale so that it is not as easy to understand them as one might expect. We believe that we do need some kind of tool to help the learner read and understand programs. Learning programs is necessary for other reasons such as program maintenance, reuse of programs. In this research, we built a model to use for program understanding. Using it, we examined possible ways of supporting program learning. Based on that, the objective of this paper is to present our efforts at developing a program learning tool, which helps narrow down the scope of a program to facilitate reading and understanding it while finding the differences of such dynamic information, as found in the program execution history, and also provide an environment in which the learner can use static information. We evaluate the efficiency of the proposed program learning tools through experiments.

Keywords: Program understanding, program learning, execution history, program reuse, program maintenance.

1 INTRODUCTION

It has been said that it is quite useful to read a program that somebody else has written. Nowadays, we can easily obtain a copy of practically any used program in our open source community. For example, we can obtain a copy of many such programs through Web sites such as SourceForge. Reading such practically used programs that are available, people can now learn high-level programming. However, the reality is that the level of such practically used programs is often complex and of such a large scale so that it is not as easy to understand them as one might expect. Furthermore, many such programs that have been open to public through the open source community might not come with the right amount of documentation. That is also a reason for people to feel difficulty in understanding programs. For that reason, we believe that we do need some kind of tool to help the learner read and understand programs. Learning programs and cultivating the ability to read programs is necessary for other reasons. For example, to do program maintenance, it is necessary to understand the programs to which some modifications might apply. Additionally, it is understood that the reuse of programs is necessary to achieve cost reduction in software development; to do that it is important to read and understand the program first to find sections of it that are reusable. In this research work, we built a model to use for program understanding. Using it, we examined possible ways of supporting program learning. Based on that, the objective of this paper is to present our efforts at developing a program learning tool, which helps narrow down the scope of a program to facilitate reading and understanding it while finding the differences of such dynamic information, as found in the program execution history, and also provide an environment in which the learner can use static information. We also evaluate the efficiency of such program learning tools using experiments.

The rest of this paper is organized as follows. In section 2, we describe the model for program understanding, and discuss obstacles preventing understanding of program. In section 3, we discuss related work. In section 4, we describe specifications of requirements for the program learning tool to be developed. In section 5, we explain how to implement the program learning tool. In section 6 we examine and evaluate how much the program learning tool we have proposed and developed can support the user and in section 7 finally we conclude the paper.

2 MODEL FOR PROGRAM UNDERSTANDING

We thought that the program understanding would proceed with three stages: assumption, checking, and recording, in that sequence, and that understanding would progress by repeating those three stages as one cycle of the understanding process (Fig. 1).

Each stage of program understanding includes the following tasks.

(1) Assumption stage

During program learning, we often encounter functionalities that we would like to understand. A program learner normally assumes a correspondence existing between the functionalities and the program that implements them. The program learner then decides what segment of the program to read based on the information of the function names, file names, and file organizations available with the program.

(2) Checking stage

This is the stage during which the program learner checks to see whether what is assumed at the assumption stage is correct or not by executing or reading the program thoroughly. The program learner would try, when necessary, to understand the program line by line and by every function in it.

(3) Recording stage

This is the stage during which the program learner would keep a record of what was checked, or in other words what was understood, during the checking stage. Through this stage, the learner would be able to refer back to what has been understood. This stage is important to progress to the next program understanding cycle.

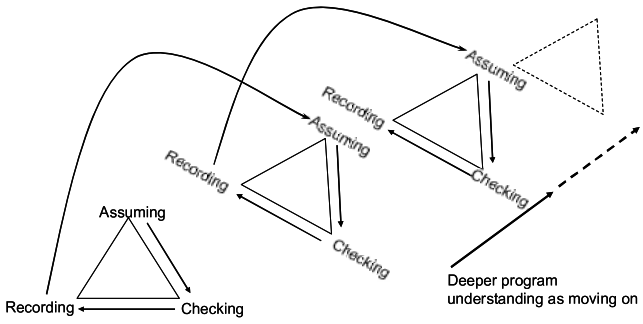


Figure 1: Program understanding model.

2.1 Information used for program understanding

The information used to understand programs is categorizable into the following two types:

• Static information

This information indicates the structure of the program to understand. It includes the information that is available from program slices and the Program Dependence Graph (PDG), which are obtained from the dependency relationship among the symbols used. With these pieces of information provided, it becomes easy for us to understand the dependency relationship among the symbols and the data used by the program and the program logic control in it. However, it would not be possible to understand programs merely using such static information: some programs contain control information, in addition to data, that a learner would not be able to know before executing the program.

• Dynamic information

This information indicates the behavior of the program to understand. It might include the information available from a debugger or a profiler. Using it, one can check to see how the program would behave at execution time so that some additional information that is not available from the static information can be provided. However, it is also necessary to know the program structure to understand it more fully. To do so, even the dynamic information would not be sufficiently good to use for program understanding.

Because those two types of information described above are mutually complementary, it is a good idea for program understanding to refer both, back and forth, during the trial and error process we would have to proceed with. Still we remain in a difficult situation to understand the practical-use level of programs.

2.2 Obstacles preventing understanding of programs

Next we examine why it is difficult to understand a practical-use level programs. To do so, we first select all the factors that we think make it difficult for us to read a program, and put them all in a list. Using that list, we can consider the reason for every one of them using a model we described previously to understand programs.

Factor 1: No good documentation is available.

During the assumption stage, we must assume that there exists a correspondence between the program and the functionalities that it incorporates. However, with no information available from the documents about the program with its functionality relationship, one would have to find it individually. In addition to the difficulty we face as to “what programs should we read?” we also have the problem that we might have to start again with a new assumption posted again if we find that our assumption was wrong (Fig. 2(a)).

Factor 2: Programming knowledge level difference between writer and reader

Readers, when wanting to learn programming, would try to read programs written from a better technical perspective than their own. Reading such a program as written by a skilled writer, a learner might notice that even a very complex process is implemented with just a few lines of code. One might fail to check to see whether the assumptions are correct or not. For that reason, it is difficult to understand the real program behavior (Fig. 2(b)).

Factor 3: Program of large scale

Using a model for program understanding, a learner might be able to understand the program gradually and more steadily while moving on with the repeated cyclic process of learning. However, it is true that the greater the number of cycles one would have to follow, the less hope one would have of understanding the program through continued reading. In addition, the learner would end up with more pieces of information to remember while progressing, which is a problem because it might cause the learner to lose understanding (Fig. 2(c)).

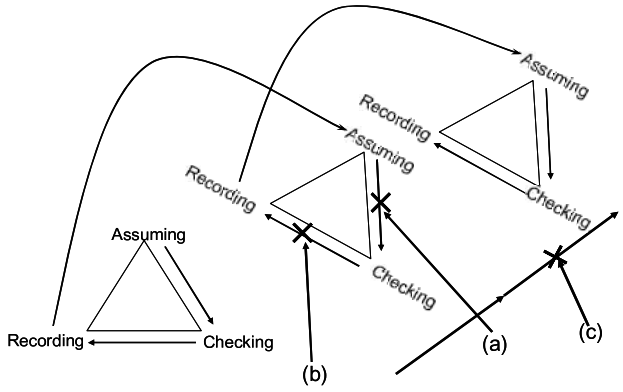


Figure 2: Reasons preventing the program from understanding.

2.3 Support for program understanding

We next consider what possible support would be available to mitigate the factors listed in the previous section, 2.2, and which prevent us from understanding programs. Regarding Factor 1, we would be able to support the learner by providing static information to elucidate the program structure. Regarding Factor 2, we believe that we would be able to support the learner by providing dynamic information to determine the actual behavior of the specific program segment. For Factor 3, we would be able to support the learner by facilitating record-keeping of information that must be remembered. However, to reduce the number of cycles used for program understanding, merely being provided with the static or dynamic information is not sufficient for support.

Wong et al. reported a study in which they extracted all necessary segments from a program based on dynamic information such as the differences existing in the program execution history. Their objective was to use that information for program maintenance [6]. Using this method, it becomes possible to read even a large-scaled program by specifically addressing the scope of what to read. This way, a learner can reduce the number of necessary cycles for learning. In addition, not only can a learner narrow the scope of program that is read based on the difference of the dynamic information such as that of the execution history. A facilitative system can help the learner understand the program by adding static information to it.

Here we describe the difference of execution history, which is the essence of our research work. First, we assume that the program we will examine contains more than one piece of functionality. The reason is that the program with a single functionality is so simple that it is too easy and easier to read than a program containing multiple functionalities. The program learner tends to want to learn one functionality first from the program.

When a program includes multiple functionalities, it is constructed with two types of parts, one of which is "the segment that is used only when a specific functionality is executed (called Proper Part hereinafter), and "another segment that is used when any functionality is executed (Common Part hereinafter). If we take the meaning of the proper part such that it is "the segment used only when a specific functionality that we want to learn is executed", then we can understand that it must be "the segment that is certainly implemented with such functionality that we want to learn".

We use two sets of execution histories when we extract such a proper part. One is created when a specific functionality that we want to learn is executed. The other is created when all other functionalities are executed. Both consist of the common part as well as the proper part. Therefore, we can extract the proper part of the functionality that we want to learn by deleting the common part of the execution history that is also created when the other functionalities are executed (Fig. 3).

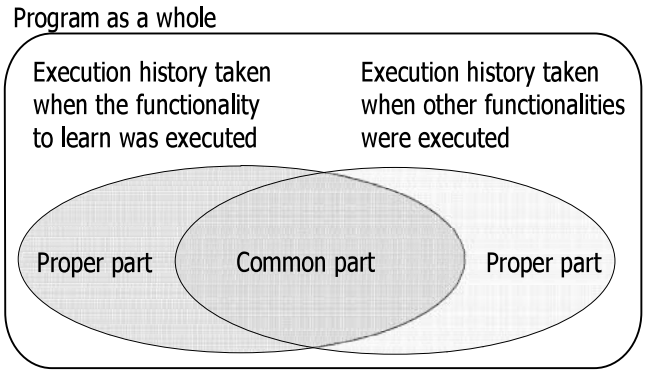


Figure 3: Differences between execution histories.

3 RELATED WORK

Ono et al. proposed a process used to help the learner understand programs and support development aiming to include the additional objective of program re-usage capability [5]. The proposal includes the process of three definitive progress stages from program understanding to program re-usage, which are of understanding the programs to change, modify, and check the possible impact resulting from modification. Through those progressive stages, they sought out what support would be required to improve efficiency. They reportedly reduced the time needed for the process from program understanding to program re-usage by 64% through introduction of techniques of both static analysis and dynamic analysis on program understanding. They established a model of the process from program understanding to program re-usage. In the present study, we concentrate on building a model for program understanding. Then we propose a tool that is derived

from the model, and concentrate then on ways to support the specific understanding of one program segment.

Knowing, as Wong et al. described, that it is difficult to locate a certain segment of a large program during maintenance, we showed that we would be able to use the differences existing in the execution histories to locate the exact segment of program to maintain as well [6]. Although the information related to the differences in the execution histories can be readily available, we can strongly support a means to extract the correct program segments. Their method was to extract the right program segments based on differences taken from multiple execution histories. Our method is an attempt at program segment extraction using differences taken from two execution histories only.

Hatori et al. analyzed list-processing programs and proposed a system to support program understanding by creating input and output samples in addition to supplying short descriptions in English [7]. They assert that their method is useful to support program understanding because their study concentrates on the scope in which input and output samples and the addition of explanation wording are created, and also because it is easy to use. It is noteworthy, however, that they are looking at list processing in which only the input and output relationship are clearly definable. However, it is probably not easy to apply it for other processing as well. Our method, on the other hand, uses execution histories, which comprise information that can always be available when any kind of program is executed. For that reason, it is applicable to various types of processing.

4 PROGRAM LEARNING TOOL

In this section, we describe specifications of requirements for the program learning tool to be developed. According to those specifications, we can describe its proposed structure.

4.1 Requirement specifications

The proposed program learning tool must satisfy the following capabilities.

- It must be able to execute the program and obtain the program execution history: the lines of code that are executed must be saved.
- It must be able to extract any difference from the two execution histories.

Based on the differences thus obtained from the execution histories, we support location of the right program segment to use to understand the program. Thereby, we can support the assumption stage, which was described above.

- It must be able to refer to both the difference information of execution histories and the static information interchangeably to proceed with program understanding.

Using the difference information placed in the execution histories, we can narrow down the scope that must be read in the program. Using the structure information of the program based on the static information, we can proceed with program understanding. This way we can support the checking stage.

- It must be able to save information that is obtained during program execution without giving any impact to the original program.

Allowing the learner to save the information the learner has understood, the learner can refer to it later when necessary. Consequently, the proposed tool supports the recording stage. It is noteworthy that if the learner would make a record by writing it directly into the program that the learner wants to understand, the program would be altered, perhaps causing some bugs that might cause a misunderstanding as a result. Therefore, it is important to be careful not to impart any alteration to the program.

4.2 Structure of the program learning tool

The following describes the structure of the program learning tool, knowing what the requirement specifications are for it (Fig. 4).

Narrowing down of the scope in the program to read supports the assumption stage, which can be accomplished by obtaining the execution histories and extracting their differences. The checking stage is supported not by the dynamic information related to the program behavior, but by the static information to give to the learner through an interface based on both the information available from the differences of execution histories and the static information extracted. Then the recording stage is supported by allowing the learner to record the information of what the learner has understood in the way a person usually attaches a sticky note on paper.

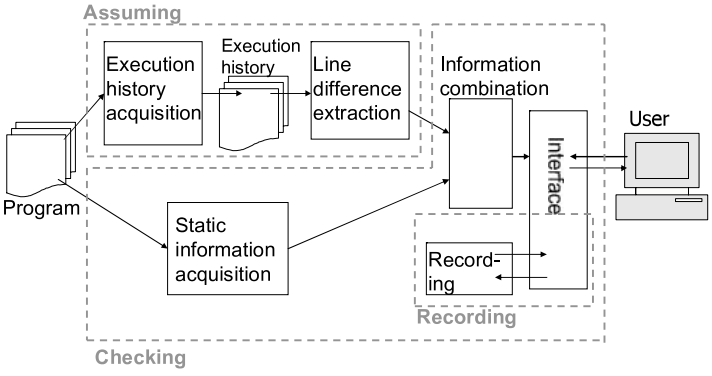


Figure 4: Program learning tool.

The following describes each function of the tool.

(1) Acquiring execution histories

This is a function to save every executed line of code as history when the program is executed. The execution history is expected to include every associated line number to save as well; duplicate lines can also be included. Every duplicate line is counted to determine how many lines were executed.

The execution history thus obtained is expected to include every unique feature of the functionalities executed. Thereby two execution histories are created, each of which corresponds to every different functionality that is executed. Taking the differences from the two histories, we will be able to find and extract the unique segments of the program.

(2) Extracting line differences of execution

This is a function to find and extract the difference of the executed lines from the two execution histories. An execution history includes the following line types.

- Lines that are executed only once
- Lines that are executed repeatedly

The information of how many times every line difference of execution is executed is quite useful, especially when the functionality we want to understand is expected to progress repeatedly based on the known program behavior, or when it is the other way around. For that reason, we allow every execution history to contain multiple duplicate lines; all such information is to be saved. Regarding the duplicate lines found in every execution history, we delete them while we are in this function of extracting line difference of execution. However, we only do so after we count them as the number of executions for the line. Subsequently, we extract the line differences from the execution histories from which the duplicate lines are deleted. The information of how many times the line appears is used by the function of combining information, as described below.

The line difference thus obtained from the execution histories represents the unique features of the program. To learn of the functionalities of a program, we would be able to narrow down the scope of the program that we want to understand based on the lines available from the line differences of the execution histories.

(3) Acquiring static information

Looking at the dependency relationship of the symbols used in the given program, we acquire static information such as the symbol dependency relationship. The symbol dependency relationship is defined as the relationship described in the following.

- Function call relationship
- Variables reference relationship
- Execution control dependency relationship

The above relationships are observed and analyzed for the symbols appearing in the program.

The obtained symbol dependency relationship is important information used to understand a program. It is also a very useful piece of information for us to use especially to find out the symbol definition location during the course of program understanding.

(4) Combining information

Next, the combined line difference information is gathered. It is obtained from the function of extracting line difference of execution and the symbol dependency relationship (static information), which is obtained from the function of acquiring static information. To view that information together with the program, we pass the combined information to the interface.

The information thus combined is provided to the learner through the interface. Then the learner can proceed with program understanding while referencing two such types of information through the interface.

(5) Recording

We receive the information from the learner through the associated interface. Then it is saved as a record. Following are three sets of information to save as a record just like we usually use to understand a program on paper.

- Information understood

The learner might want to describe some information understood so far and leave it in a place near the associated location of the program when reading a program on a paper basis. Similarly, such understood information is recorded using this recording function while finding the correspondence between what was understood and where it is located in the program.

- Information of category and importance level

When reading a program on paper, the learner might want to highlight it or put a colored sticky note to a specific location of the program to categorize it as what has been understood or to show the importance level. Similarly, categorization and the importance of program segments are visualized using this recording function.

Using this function, the learner can easily make a record of what has been understood in the program. Furthermore, what has been understood can be available and referenced whenever necessary.

(6) Interface

This is a function that provides the learner with the recording function in addition to the information obtained using the function of combining information in a way that the learner can reach without much difficulty.

5 IMPLEMENTATION

This section describes how to implement the program learning tool that we are proposing. In practice, we implemented it under such an environment as that shown in Table 1.

Table 1: Implementation environment

OS	Fedora core (Linux kernel 2.6.17)
CPU	Intel Xeon 1.66GHz * 2
Memory	2GB

Figure 5 depicts the implementation of the program learning tool. Below we will describe each module of the program learning tool.

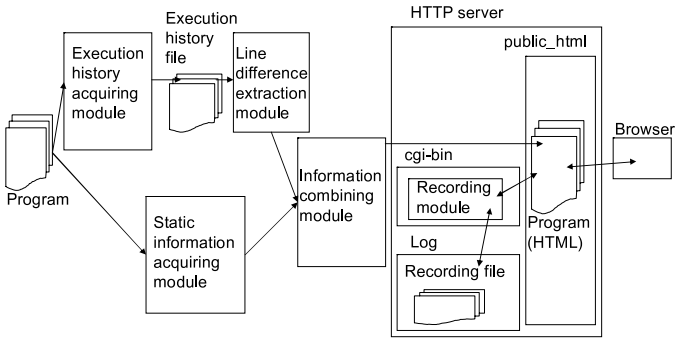


Figure 5: Implementation of program learning tool.

(1) Execution history acquiring module

This module executes the given program and writes the history of every line execution to a file. We allow the execution history file to be written out file-by-file for every file as long as it includes all the executed lines. Additionally, we allow duplicate lines to be included in it. It is implemented in a form in which we modify the famous debugger, gdb-6.6 [8], to obtain an execution history properly.

Inside the gdb, when the program passed is executed step-by-step, the print_source_lines_base function, which exists in gdb/source.c, is called. This function holds the executed lines as an argument to be executed step-by-step. Knowing this fact, we modified the function to be able to output the executed lines to a file by delimiting every executed line using a carriage return character. Therefore, we modified that function to acquire an execution history. Furthermore, we obtained every execution history written out to the file that included the associated executed lines. We set the naming rule as follows for the file to which execution history is written.

Naming rule: (dirname)%(filename).trace
Note: "/" if found in dirname must be replaced by "%".

(2) Line difference extraction module

From the execution history file group, which is obtained when one of the two different kinds of functionalities is executed, the module described in this section extracts the executed line differences from every such execution history file, and outputs them to a new file. Duplicate line numbers, if found in any execution history file, will be deleted, but they will be counted as they have appeared; that information is written to a file together with the line differences. We will use a shell script and perl because they make it easy to use and manipulate files and character strings.

To have easy manipulation for every execution history file group, we will put the same group files under one directory. Then from the execution history files, each of which is placed in one of the two directories, we will extract the line differences in the following process. First we define the directory A, which contains the execution history files obtained when the functionality the learner wants to understand is executed and another directory B, which contains another set of execution history files obtained when all the other functionalities are executed.

1. Find and get an execution history file a from A, and the same name file a' from B

- If found, we will

(a) Create a hash table,

(b) Define the line difference for such a line that is in a, but not in a', and

(c) Using the line difference as a key to the hash table, count up by one on the appearance count entry of the table, which the key points, and

(d) Repeat the above process from (a) to (c) for every line written in the file, a.

- If not found, we will

(a) Create a hash table,

(b) Define the line difference for all the execution history lines in the file a, and

(c) Using the line difference as a key to the hash table, count up by one on the appearance count entry of the table, which the key points, and

(d) Repeat the above process from (a) to (c) for every line written in the file a.

2. Write to the file both the line differences, each of which is a key to the hash table and the appearance count that every key points.

(3) Static information acquiring module

This module examines and analyzes the given program. It obtains the relationship among the symbols used in the program. Here, what is analyzed will be the relationship of function calls and the variable references. We will use global [9] with options F, a, and n here because it is easy to implement and enables analysis of the symbol dependency relationship and to output the program in an HTML format with the dependency relationship used as a link.

- Option, F
  Used to put the contents partially into the frame.
- Option, a
  Used to add alphabet indices.
- Option, n
  Used to display line numbers.

(4) Information combining module

This module allows visualization of the resultant line difference information together with the program that is placed in an HTML format.

We will use a shell script and perl because they make it easy to use and manipulate files and character strings, and place the line difference information in an HTML format. To allow the learner to refer to the line difference information together, the module is implemented by changing the original source code in HTML format using awk.

Consequently, the rewritten HTML file has a description telling where each line difference exists in terms of which line of what file. Using that rewritten file, we can refer, using the global, to the corresponding program that has been rewritten into an HTML format based on the information of the line number or the file name given.

(5) Recording module

This module receives information from the program that the learner wants to record and save as a record. To do so, we must be sufficiently careful not to give any impact to the program by writing the information of what was understood directly into the program.

We managed the sticky notes using javascript. We implemented this module to be able to record the information through the cgi that was created using perl. To make it functional with the program placed in HTML, we change the program in the HTML format using awk.

The recording module saves the following types of data by writing them in a recording file in a CVS format. Furthermore, each recording file is created for every associated file that is changed to HTML format.

(id),(x-coordinate of sticky note),(y-coordinate of sticky note),(sticky note size on x),(sticky note size on y),(color of sticky note),(character string)

Now that we are using HTML, the learner needs only a Web browser to use the system. Figure 6 shows the operation screen of the tool.

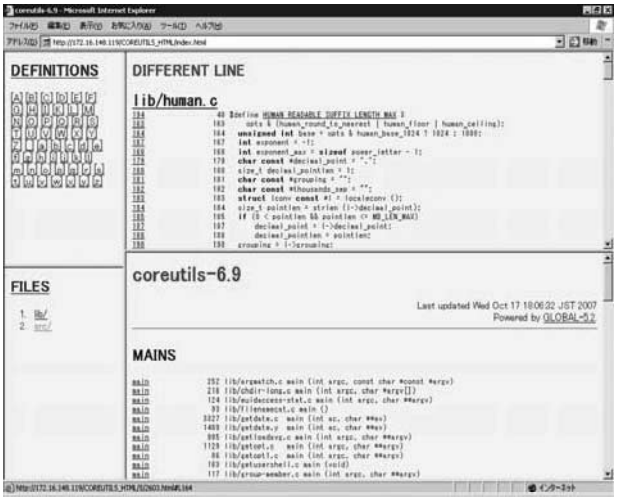


Figure 6: Operation screen of program learning tool.

6 EVALUATIONS

Next, we examine and evaluate how much the program learning tool we have proposed and developed can support the three stages of the program understanding model.

6.1 Evaluation to assumption stage

We will first evaluate how much we can narrow down the scope of program that the learner must read based on the differences found in the execution history of the program. If we can prove that we can narrow down the scope to read, it would be a good means of supporting understanding of programs, even for large-scale programs, for which it is said that no good support is currently available.

6.1.1 Evaluation method

We will use coreutils-6.9 for evaluation purpose, which is sufficiently large and which includes UNIX commands (Table 2).

Table 2: Structure of coreutils-6.9

Total number of directories	16
Total number of files	1173

We will look at the ls command and the wc command of coreutils to evaluate the extent to which we can narrow down the scope to read based on the differences found in the execution histories. Regarding those differences, we will use the following.

- The differences that can be found when we have an option to the ls command and when we have no options to the ls command
- The differences that can be found when we have an option to the wc command and when we have no options to the wc command

6.1.2 Evaluation results

The results we obtained for the difference we had when we had an option to the ls command and when we had no options to the ls command are presented in Table 3.

Table 3: Differences between execution histories for ls -l and ls

ls (lines/total)	ls -l (lines/total)	Differences	Difference ratio
423/7003	581/9533	351/9533	3.6%

In addition, Table 4 shows results obtained for the difference when we put an option and the argument of 5-

line and 5-character text file to the wc command and when we had no options to the wc command.

In light of this experiment, we can recognize that the number of lines extracted as line differences for both ls and wc are very small compared to the total number of lines of the program. This result shows that we can narrow down the scope of a program for the learner to read and understand. Thereby, the tool we developed can be very useful to facilitate understanding, even for a large-scale program.

Table 4: Differences between execution histories for wc -l and wc

wc (lines/total)	wc -l (lines/total)	Differences	Difference ratio
90/1149	99/1227	17/1227	1.3%

6.2 Evaluation to checking stage

We will measure the time spent to understand a program of coreutils, whose information is given in Table 2, and compare it for the two cases: one in which we use the program learning tool we are proposing and another in which we use an editor and necessary commands only. We can expect that the time spent with the program learning tool is shorter than the time spent using an editor and commands only. If that is the case, we can tell that we will be able to use it as a good tool to support the checking stage of the program understanding model.

6.2.1 Evaluation method

We will measure the time to spend for four different types of task that will have to be done to understand a given program (Table 5). We will then compare the work duration time when using the program learning tool and when using an editor and commands only. Six undergraduate students volunteered to participate in this experiment.

Table 5: Task description

A	Search for the location of a function definition
B	Search for where a function is referenced from
C	Search for the location of a symbol definition
D	Search for where a symbol is referenced from

6.2.2 Evaluation result

Figure 7 shows the average work duration time for the students to perform tasks A–D using the program learning tool or using an editor and commands only. With an editor and commands used, we can see that it includes the time to open and close the editor and to execute the commands. With the program learning tool used, however, the same work can be accomplished by merely tracing the associated link. Consequently, the work time spent while using the program learning tool was less than that using an editor and commands only. It is noteworthy, however, that slightly more time was spent than expected to do task C even with the tool used. Why it happened was because the static information acquisition module, global, which we used this time, failed to catch the symbol dependency relationship perfectly. For that reason, there existed some left-over symbols that were not linked properly. Despite that fact, we believe that we can shorten the work time for program understanding: we can support the checking stage as well.

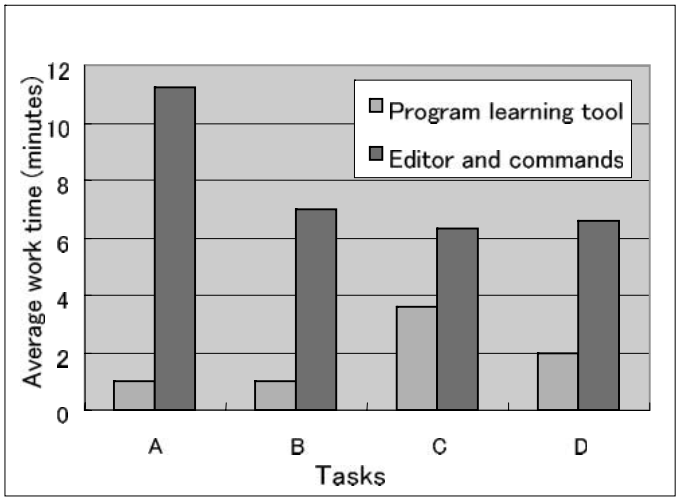


Figure 7: Comparison of average work time for each task.

6.3 Evaluation of the recording stage

We distributed a questionnaire; we evaluated the responses related to the recorded information and how to use the recorded information. If the responses would have reported that it was insufficient, we would need to resolve the issue in the future, and improve the tool to support the recording stage in a much better way.

The questionnaire requested an explanation of how the program learning tool would support the recording stage of program understanding in addition to what information should be recorded and what functionalities would be necessary to use the recorded information. The questionnaire was given to the same undergraduate students as before. The forms were designed to allow the participants to fill it out in a descriptive fashion; we expected frank opinions to be reported.

6.3.2 Evaluation result

Following are some typical responses we received.

・ Would like to see the list of most recent sticky note information.

- Would like to search for every specific sticky note information.
- Would like to change the color not only of the sticky note, but also of the characters on them.

Based on the valuable opinions we received, we think further improvement is necessary, especially on the following items.

- Expansion of sticky note management capability

Emplacing sticky notes physically while reading a program on paper, it is true that we can visualize and understand instantly where every sticky note is placed. However, we cannot do the same with our program learning tool. Consequently, we think expansion of the function of the tool, for example, would be useful to be able to manage the time at which every sticky note is placed, so that we can notify the learner as to when the most recent one was placed.

- Manage the recorded information in a database

When we read a program while emplacing sticky notes, the problem might be that the more sticky notes we place, the lower the chance that we can find the right information. We can expect the same situation to occur as a greater amount of sticky note information is being added using our program learning tool. We think we would need some function to search for the right information that has been recorded to solve this problem. To do so, a database to manage it would use the CVS format, which we are doing at this moment.

- Improve the usage

When reading a program on paper while emplacing sticky notes on it, the learner determines how to handle the sticky notes and how to put comments on them. The program learning tool we developed is not as flexible as that; for better usage in the future, we intend to improve and expand its function.

7 CONCLUSION

Narrowing down the scope that must be read helps the learner reduce the number of cycles of the program understanding model; consequently, it reduces the chance of the learner losing the willingness to continue the process of program understanding. In addition, using the static information that represents the program structural information, the checking stage of the program understanding model can be supported. Moreover, it enables the learner to record the information of what has been understood just as one might physically leave a message on a sticky note for later reference, which helps prevent the learner from losing track of what has been understood.

In our research work, we developed a program learning tool that can support the learner to understand programs based on the program understanding model. Using all of these functionalities provided, we expect that the cost for program understanding can be reduced significantly. At present, however, the following issues remain to be solved.

- Using the program learning tool we developed, no dynamic information for how the program behaves is used during the checking stage. Under these circumstances, undefined data cannot be processed without executing the program. Some improvement to the tool must be made by considering the dynamic information of the program execution behavior.
- Because the interface of the program learning tool we developed uses HTML, it can support more than one learner for program understanding concurrently on the Web if we were able to improve the tool further. To do so, however, would require such information to provide in the recorded information as “who put it down” or “whether we should make the sticky note information open to the public or not”. These are issues that must be considered to determine whether to expand the learning tool function.

REFERENCES

[1] Diomidis Spinellis, “Code Reading: The Open Source Perspective (Effective Software Development)”, Addison-Wesley Pub (Sd) (2003).

[2] The Apache Software Foundation, http://www.apache.org/, 2007/08/19

[3] eclipse, http://www.eclipse.org/, 2007/06/22

[4] SourceForge.net, http://sourceforge.net/, 2007/06/13

[5] N. Ono, A. Honma, T. Fukaya, ” Proposal of a Program Understanding/Development Process with Mutual Complement of the Static-Analysis and Dynamic-Analysis”, Technical report of IEICE. KBSE, Vol. 100, No. 441, pp. 9-16 (2000).

[6] W. Eric Wong, Swapna S. Gokhale, Joseph R. Horgan, ``Locating Program Features using Execution Slices", Proceeding of the 1999 IEEE Symposium on application – Specific System and Software Engineering and Technology, p. 194 (1999).

[7] K. Hatori, M. Nagata, ” A Support System for Understanding List Processing Programs Which Generates Explanations and Input-Output Examples”, The transactions of IEICE, D-1, Information and System, I-Computer, Vol. J77-D-1, No. 6, pp. 443-453 (1994).

[8] GDB: The GNU Project Debugger, http://sourceware.org/gdb/, 2007/06/07

[9] GNU GLOBAL source code tag system, http://www.gnu.org/software/global/global.html, 2007/11/05

[10] linux kernel hack japan, http://hira.main.jp/wiki/pukiwiki.php

[11] GCC Reading Rm., http://wikiwiki.jp/aloha/

[12] Read Tcl 2.1, http://tcl.s27.xrea.com/mains.html

[13] PostgreSQL Reading Rm., http://testnoda.sakura.ne.jp/
[14] Y. Tanaka, T. Hagino, ” Introduction of social navigation in comprehension of source code”, IPSJ SIG Notes, Vol.2007, No.12, pp. 85-90, 2007-CE-88-(12) , (2007).
[15] gonzui: a source code search engine, http://gonzui.sourceforge.net/, 2007/12/5.
[16] Y. Sawada, H. Ohkubo, H. Kasuya, S. Yamamoto, ” A Software Browser for supporting communication with sticky notes”, Technical report of IEICE. SS, Vol. 103, No. 189, pp. 13-18 (2003).

Considering Leadership in Education

Kouji Yoshida*, Isao Miyazi**, Kunihiro Yamada***

*Graduate School of Electrical Engineering, Shonan Institute of Technology, Japan
**Faculty of Informatics, Okayama University of Science, Japan
***Professional Graduate School of Embedded Technology, Tokai University, Japan
yoshidak@info.shonan-it.ac.jp, miyaji@mis.ous.ac.jp, yamadaku@tokai.ac.jp

Abstract With prior experience of working in the industry, I have taught students at university for the past eight years. During that time, I have specifically focused on training people in leadership and have investigated ways to accomplish that task. My research work has shown that training people for leadership necessitates sustained enthusiasm and continual instruction, which also applies to leadership training in the educational world. An important accomplishment as a leader is having every person set an appropriate personal target and then raising that target at appropriate times. Based on study results, along with classical and recent modes of leadership, this report describes the concept and practice of leadership training for today's education world.

Keywords: leadership, education, motivation, learning, training

1 INTRODUCTION

After working in the development of information systems for use in industry, I trained engineers and first-year employees at a company for 14 years. In addition to this employment, I also worked as a part-time instructor at universities and technical junior colleges over 10 years. During that period, while working for the company, I spent 3 years completing a doctoral course at graduate school. Thereby gaining valuable experience and knowledge. After acquiring the doctorate, I taught students at university for 8 years, during which time I learned simultaneously that teaching was not an easy job and that learning was refreshing.

During my work at the company, I engaged in employee education and examined how information technology education and engineer training should be conducted. Consequently, I considered not only how to set up a curriculum, but also how to train Information Technology (IT) Engineers [(1)–43].

This paper discusses leadership models based on such prior experience at a company and a university. While referencing some recent theories related to leadership, I formulated a practical means of fostering leadership: what to do, what to care about, and what to acquire.

2 EXCERTS FROM TOPICS ON LEADERSHIP

The following are some reports from relevant literature available on leadership.

- Professor emeritus Argyris of Harvard University talks about defensive reasoning[(5)]:

An active, overly enthusiastic teacher might think that the learning in a class, which was not progressing as well as expected, was his fault. He might try to be more enthusiastic, but his students might then be rather depressed. It is important that the leader and the followers share a good relationship with good communication passed back and forth between them.

In general, the belief that the students should be as enthusiastic as the teacher is outdated. It is important, however, that good communication be established between the person who teaches and those who are taught.

As Isoroku Yamamoto has said, "People will not follow you unless you first show them how to do it, explain it next, let each do it, and praise them." William Arthur states boldly: "The mediocre teacher tells. The good teacher explains. The superior teacher demonstrates. The great teacher inspires." It is important to make students enthusiastic. A good teacher can do this.

- From "The Conditions of Management" by Peter. F.Drucker [(6)]

How is a capable person expected to behave? What should the executive do in these modern times? During the era of mass production, there were executives who posed no problem (basically, employers were also positioned as executives). They were few and their roles were usually structured in a hierarchy. Today we are surrounded by executives––financial traders, engineers, and management staff—who all work with some type of

information. How should they be evaluated? How can their productivity be improved? To do so, Drucker describes how we should establish an environment in which such workers can make a habit of working with the skills they have acquired to improve their performance effectively.

Additionally, Drucker reasonably describes that sufficient intelligence is not necessarily the factor engendering their higher performance. Knowledge, he describes, might be hazardous for one reason or another. He explains further that, "The intelligence, imagination and knowledge are nothing but very fundamental characteristics. To produce the final outcome, you would need the right ability to achieve high performance." he goes on to say that "The ability to achieve high performance can be acquired."

In short, it is true that one needs knowledge and intelligence, but they are not sufficient enough. One must learn to acquire skills and abilities to achieve high performance repeatedly, and make it a rule to do so. Otherwise, an executive will not be able to assume a role among human resources that can contribute to high performance of a group.

• John P Kotter's Leadership Theory[7]

Kotter describes that leadership must transform itself concomitantly with change. He describes, "What can contribute to achieving change the most are links through an informal type of human resource network. The formal structure of hierarchy is useful only in a stable environment." He therefore asserts that it is important to retain informal relationships with others in everyday life.

• Gerald Weinberg's Theory of Training Engineers[8]

By examining system developments that have produced success and those that have not, we can understand almost all of the former depend on the contributions made by a small number of outstanding engineers. Such super engineers would need from now not only to have the ability to organize and motivate others, but also to change technology. What he describes is that they are required to have a "problem-solving type of leadership." A "carrot and the stick" type of leadership might work for simple work to be done by many people in a short time. However, it might not often work out in a problem-solving type of situation in which one builds a computer system or develops new technology. In this environment, one must have an "organized" perspective, along with the patience to wait and allow each person to grow as a human being.

• K. Anders Ericsson's Expertise and Expert Performance[9]

Expertise and experts can be developed through repetitive efforts. Additionally required are a good teacher and a good timing of development. A good leader knows what problems to give and at what timing to have workers accelerate their learning speed and develop their skill level to the next step. The expertise and the expert are developed with three factors: practice, opportunity, and luck.

In short, leadership is a product of learning. Leadership is something that a person learns to acquire. Furthermore, it is acquired as a skill along with rules for daily life. Therefore, leadership training requires repetitive, effective practice.

• Patrick J. McKenna's Coaching Theory [10]

McKenna wrote a book for those who want to be a leader for the first time, who must assume the role of a leader and as a member of a group, answering the question of how to be a leader before all professional members.

The leader's role differs fundamentally from that of each participant. A leader must be a person who can initiate something new by guiding people in terms of their feelings, either in an office environment or in a classroom. A leader is expected to deliver energy and hope. The leader's role is to heighten their people's eagerness and enthusiasm. Furthermore, a leader must give rise to a consensus among team members.

To achieve this, what does a leader have to do for their team members?

○ What a leader can do for their group is:

(a) Have his members create energy and eagerness;
(b) Be a source of creative ideas to stimulate creativity from his members;
(c) Forge teamwork;
(d) Help develop a common goal that everyone can work towards;
(e) Help team members solve problems and remove troubles, if any, to support team members to engender their success;
(f) Advise and exchange opinions among them to help them think about their own problems thoroughly;
(g) Have them follow the standard rules (if, however, somebody does not comply with it, act promptly and firmly, but gently);
(h) Be a "conscience" of the group when self-discipline fails (giving pressure gently); and
(i) Continue encouraging them to improve their performance, quality, and efficiency.

What is the difference between a leader and manager? A manager is a person who must complete routine work continually without fail. A leader, on the other hand, is expected not only to do that, but also to look to the future and be flexible as the team environment changes and as the outside world changes. That is the greatest difference between them. The world appears to be changing slowly and gradually, but one would note great changes 5 years or 10 years from now. That difference would constitute a large problem if it were not recognized. The team would lag far behind in coping with the new changed environment.

3 METHODS AND PRACTICE OF LEADERSHIP TRAINING IN EDUCATION

The method and practice of leadership through classes and seminars at school is described below.
(1) Teaching a class

- Preparation

Preparation is important in any conceivable activity. Explanation in advance as to what is to be taught next enables students to prepare themselves well and be ready to go.

First some authority in the class is necessary. Then the instructor should address students on how to prepare for the lectures and how to receive the classes. The instructor must compel them to understand the importance of concentration and continual effort. The instructor must teach the class with the approach that this is a class with many good students. The instructor must also praise them if they behave well. Thereby, the students and instructor come to enjoy a shared atmosphere.

- Practice of teaching a class

My first teaching experience was a class of first-year students in the department with a lecture on "Document creation on information technology." The class was taught while I, the author was investigating the subject of leadership. The students were able to improve themselves as described below.

I tried to raise the target level and the problems given to them gradually just as one would dangle a carrot in front of a horse. In addition, I gave them mainly repetitive exercises. During that process, I gradually raised their level of capability. In doing so, the students were able to sense, independently, that their abilities and skills were improving. Thereby becoming more motivated.

· Example of a report writing class

First, I explained how to write a report, telling them the main idea of the class. Subsequently, I gave them a report writing exercise, but they were unable to do it as I had expected. The problem was that they were unable to write good sentences. What I did then was to have them practice writing down what each had understood so far. Figure 1 depicts the relationship between recognition and understanding of oneself and their expression. It is not difficult to express what you understand, but it is quite difficult to write about what you have read.

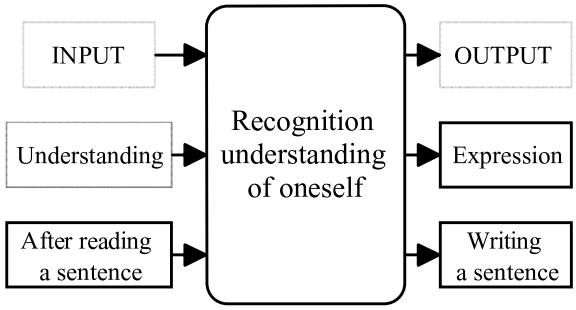


Figure 1: Relationship between the recognition understanding of oneself and expression

For example, imagine that you had to write sentences explaining a bicycle to somebody who does not know about bicycles. First one needs the ability to express what one has understood. Even if you were to understand it well, you would know that you might have trouble passing the correct information on to another unless you could clearly express yourself. The next exercise is one in which a student reads the information provided and then summarizes what they have read.

The exercise is one of summarizing an extract from a book, and expressing it in their own words. At every step of the exercise, I guided the students by showing a good example of a report with some comments added. I advised them to do the same the next time.

Another exercise required the student to summarize a passage, write an outline of the contents of the description of how to write a scientific or medical report. After repeating exercises of this type, the student was able to understand more clearly how to write a report. The students had improved their report writing skills. Such skills must then be improved gradually in one way or another in a spiral fashion. A point to note is that anyone would tire of such repetition, continuation would be impossible without some

content changes. That is a difficulty that must be confronted by the instructor.

・Necessity of repetition

An instructor often encounters the situation in which students must remember information or act in a certain manner.

To do so, they would merely need to repeat the information to themselves, as others would do similarly. One would be unable to learn to swim if they were told only verbally. The same is true for operating a Personal Computer (PC). Physical practice is necessary.

To master operation of a PC or manipulation of computer programs, one would need to practice repeatedly rather than just think about it. Table 1 shows how to master Technical Documentation, Learning English, and Leadership.

Table 1 Comparison of Technical Documentation, Learning English and Leadership

Item	Technical Documentation	Learning English	Leadership	Remarks
Basic acquisition	Learning a typical pattern	From sound, word, phrase to sentence, managing a fixed form pattern	Learning a model of typical leadership and a corresponden ce pattern	Used other than learning
Application	Put an opinion of self and repeat	Apply a fixed form pattern to conversatio n	Learning while in a real workplace	Unconsciously
Intellect	△ Cerebrum nucleus	○	△	Cerebrum
Physical strength	○	△	○	Cerebellum
Response	△	△ Speech center	◎ Motor center	Spinal cord, Cerebellum
◎deeply concerned, ○considerably concerned, △concerned				

Writing a technical document first requires a concept: construction of the idea and anything abstract related to the contents. In contrast, mastering conversation and listening in English is relies entirely on how much one practices. The same is true for computer programming. Acquiring leadership, however, is impossible, even with as much knowledge as one can obtain, until one practices sufficiently. In that sense, it is important to train for everything using both the cerebrum and cerebellum in a balanced manner.

(2) Seminar classes

An effective learning method is to use comparisons with other members of a group, teaching how to motivate and how to apply. It is effective having motivation through competition. It is hard for a professor alone to give instruction. If a student does not understand something well, it is rather important to make the student notice the current situation of the team, and the status of other team members. Another effective way is to apply what is presented in Figure 2, Group dynamics of learning, in which one leads the teams in the right direction depending on the situation. That is the state of "flux" from which you can easily make a change. Once a drawing takes place, the state becomes stable.

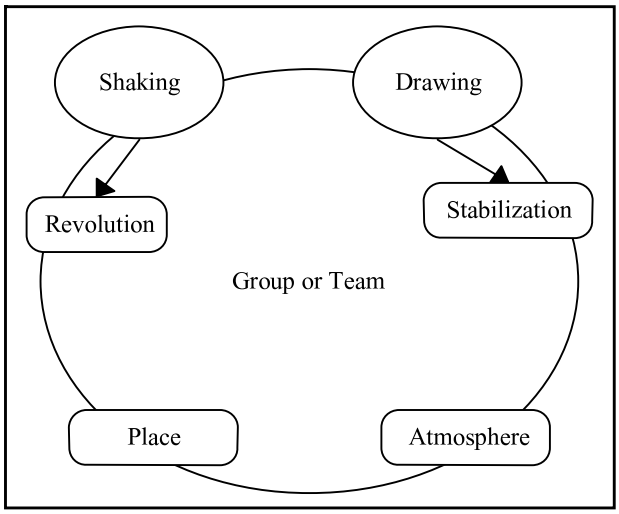


Figure 2: Group dynamics of learning

・A guide for seminar class students on daily life behavior

It is important is to have good habits in daily life. It might be necessary to start by changing daily life. For example, a student must not come late for school. A student must be punctual and always on time, and must finish work before the deadline. No such habits can be acquired easily by hearing a warning once or twice. Repeated daily guidance is necessary until a student can develop such habits without having the feeling that it is strange to do so.

In addition, students should be involved in cleaning the study area or class room. Such a duty might be assigned for every student to do once a week: for example, the student would throw out the trash from the seminar room, Once such a task is assigned and students are allowed to think about it and do it independently, it works much better than might be expected.

The student with no prior habit of putting things neatly and cleaning a personal room in everyday life

would start doing it under this training. It might be the starting point of a whole new way of life. In doing so, the student's brain is also going to form in the habit of being more orderly, resulting in leading everything else to be in a more organized state.

- A way to have students acquire leadership

During the seminar session, assigning every student a role ie; leader, sub-leader, accountant, or party caretaker, making everyone take responsibility for something, is a good way to help students become more independent.

Additionally feedback to other team members is always important in the seminar class because even a good student will feel ignored and lose enthusiasm once communication is lost. The importance of feedback is depicted in Figure 3. The leader should take good care of those students who tend to have insufficient communication with others and avoid it. The leader might be advised to make a checklist to cover that. The leader should understand such characteristics of students to maintain students' enthusiasm[11].

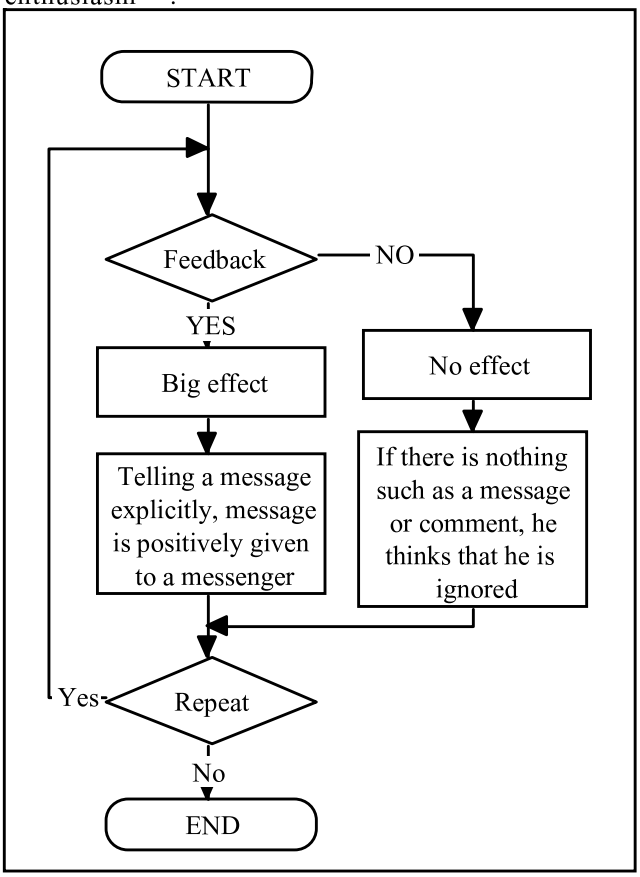


Figure 3: Importance of feedback

4 COMMUNICATION

Encouraging students to have good communication amongst themselves is not as easy as one might think. Everybody should understand how important communication is. As Weinberg observed, students do not notice errors or success by themselves. To assist them, a leader must watch them and give feedback. From other's opinions and the information that you happen to record, you are often able to get the correct information to use.

On the other hand, it is true that a manager will tend to think based on the state in which he or she is in. One might think that all of what was explained has been understood, but a communication gap exists. Figure 4 presents the form of recognition of oneself: a person might be merely reflecting another's opinion, absorbing without understanding, refracting the information, or otherwise filtering it.

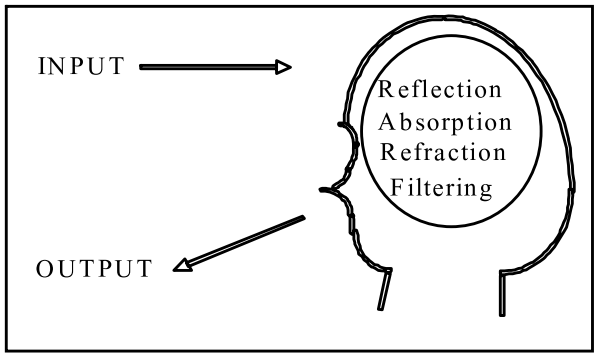


Figure 4: Form of recognition of oneself

There should be no problem if both parties recognize in each other that you are already have good communication.

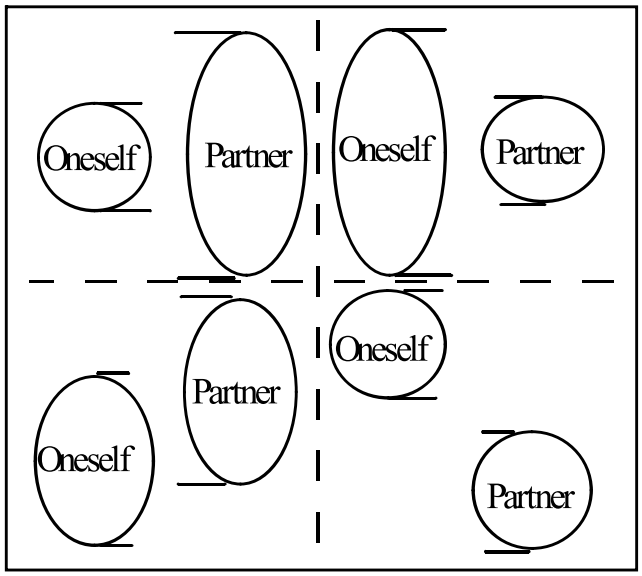


Figure 5: Pattern of mutual recognition

However, as shown in the bottom right corner of Figure 5, even given a pattern of mutual recognition, if two people, say an instructor and student, have no common information to discuss with one another, no communication can be established.

5 CONSIDERATIONS

Leadership is something that others feel when one is doing it. The key to motivating others to do as the leader expects is praise, appreciation, forward thinking, assistance and support. However, on some occasions encouragement or even reproach is necessary.

How do people learn? Fundamentally, it is up to their level of enthusiasm. Consequently, what can we do to raise their enthusiasm? Mutual understanding is necessary. A lecture or seminar class is merely the start. The starting point might be different for different people. For that reason, one would need to establish various opportunities.

Forcing students would not work. It is said that you can lead a horse to water, but you cannot make it drink. The right situation would have to be actively constructed, or patience would be needed while the student gets ready to take action independently. However, merely waiting forever is not the right answer. Therefore, the setup of the appropriate situation is an important and often necessary task.

6 SUMMARY

Education and leadership are fundamentally person-to-person relationships: they pertain to a connection between a leader and their followers. The basic idea behind the theory of true leadership should be the same when it comes to leadership in the education world. Mutual recognition must be established through good communication to generate the appropriate shared awareness. It often happens that one person understands something very differently from what is expected. Often, students don't seem to understand "the meaning of the problem given" because they do not understand the basic concepts and tasks, even though I have no doubt about their understanding of my expectations.

An orchestral conductor must tell their players similar things during day-to-day practice: what to do and not to do as well as their philosophy about the piece. The real performance of the music is the final outcome from all such efforts that the conductors and performers have made. The atmosphere at that time is expected to determine the excitement created between the music and the listeners.

Similarly, the lecture and the guidance by a leader are also a live performance. They can be applauded on a live image with others through right communications that are pursued with others. Others are ready to accept that role[12].

During class lessons or seminar classes in the education world, the atmosphere at every moment are important, in addition to the atmosphere that the lecturer and the students created together on-site. The relationship between the leader and the followers is also crucial. Even if the leader is active and powerful, the followers might not come together. The receivers must be as powerful as the leader is. Therefore, what is important is first to know of the knowledge level and the energy level of others. Based on that, a lecture can be planned accordingly. Finally, a teacher learns while teaching, which should be kept in mind when giving great effort on a day-to-day basis. Remembering that fact, and reflecting on its importance, is necessary for success in the long run.

This study was supported by a government grant for scientific research spending, #18500737.

REFERENCES

[1] Kouji YOSHIDA, Masahiro KURODA, Hiroshi ICHIMURA,T adanori MIZUNO, Sanshiro SAKAI, "A Study Of Next Information System Education" IPSJ, Vol.2001, No.3, pp.17-23(2001).

[2] Kouji YOSHIDA, Masahiro KURODA, Hiroshi ICHIMURA, Tadanori MIZUNO, Sanshiro SAKAI, "A Study Of Next Information System Education" IPSJ, Vol.2001, No.3, pp.17-23(2001).

[3] Kazuhiro Nakada, Tomonori Akutsu, Chris Walton,S atoru Fujii, Hiroshi Ichimura, Kouji Yoshida, "Practice of Linux Lesson in Blended Learning" ,KES'2004,Proceedings,Part2,Knowledge-Based Intelligent Information and Engineering Systems Part2, pp.920-927(2004).

[4] Kouji Yoshida, Isao Miyaji, Kunihiro Yamada and Hiroshi Ichimura,"Distance Learning System for Programming and Software Engineering", Knowledge-Based Intelligent Information and Engineering Systems: KES 2007-WIRN2007, pp.460-468.

[5] Kanai Toshihiro, "An introduction to leadership", Nihon Keizai Shimbun, Inc. (2005/3)

[6] Peter F. Drucker: "The Effective Executive", Harper Collins(1967)

[7] John P.Kotter : "ON WHAT LEADERS REALLY DO", President and Fellows of Harvard College(1999)

[8] Gerald Weinberg, "Becoming a Technical Leader: An Organic Problem-Solving Approach" , Dorset House (1986/09)

[9] K.Anders Ericsson, Michael J. Prietuula, Edward T. Cokely,"The Making of an

Expert",Harvard Business Review (2008/3)
[10] Patrick J. McKenna , "David H. MaisterFirst Among Equals: How To Manage A Group Of Professionals", Free Pr (2005/4/4)
[11] J.Sterling Livingston"Pygmalion Management", President and Fellows of Harvard College(2003/4)
[12] Harvey Seifter, Peter Econom, "Leadership Ensemble(Lessons in Collaborative Management from the World's Only Conductorless Orchestra)", Orpheus Chamber Orchestra(2001)

Supporting tool for student who learns use case modeling

Teruhisa Ichikawa[†], Chikage Ohta[††]

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

Abstract: Increasing numbers of students of faculties related to information technology are learning UML (Unified Modeling Language). In order to effectively model the object area, it is necessary to understand about business. It is difficult, however, for the student who doesn't know business to call up the tacit function not described in the requirements specification. This thesis proposes to use ontology as a way to draw out a tacit function when the student makes the use case diagram, and is a verification of the effectiveness of the procedure.

Keywords: UML modeling, use case diagram, ontology

1. INSTRUCTION

Increasing numbers of students of faculties of informatics are learning UML (Unified Modeling Language). In order to effectively model the object area, it is necessary to understand about business. It is difficult, however, for the student who doesn't know the business to call up the tacit function not described in the requirements specification. Moreover, it is necessary to learn the modeling technology in a limited time.

In this thesis, we propose the study support tool for the use case modeling that requires business knowledge.

Use case modeling is used to extract a necessary function (it's called use case) from the requirements specification such as RFP (required for proposal), and to make a model. The model refers to the ontology of a limited object domain. The result is described as a use case diagram. The tacit function, i.e. knowledge not described in the requirements specification, can be discovered by referring to the ontology of the object domain.

To grasp the concept of ontology and to discover tacit knowledge of an object area, and to identify requirements specification, it will be useful for students to model a familiar object area, such as a library or convenience store.

In this thesis, we aim to develop educational support for making the use case diagram by selecting an object domain likely to be familiar to students, enabling them to make a more refined use case diagram.

2. REVIEW OF CURRENT RESEACH

Current research that applies ontology to UML includes the following:

Minegishi's "Supporting Software Engineering Processes with Ontologies"[1] is object modeling research aimed to support making the analysis class diagram that is the product in the systems analysis phase. It proposes to solve the following three problems:

(1) Modeling cannot be done due lack of understanding concerning business.
(2) System specific information and common domain information cannot be separated.
(3) Significant differences of interpretation pertinent to the product are generated.

To solve these problems the analysis class diagram making a support technique using ontology is proposed. The process of making the analysis class diagram is supported by the main noun extracted from the use case description which refers to the domain ontology and general ontology that systematizes the concept in the domain.

Kamiya's "Supporting Analysis Class Modeling with Ontologies" [2] is concerned with object modeling too. If the domain ontology is insufficient, an acceptable class diagram can be derived by applying general ontology, combined with conversation processing with the modeler, etc.

To date no research of support for making the use case diagram from the requirements specification proposed in this thesis is available.

3. OUTLINE OF ONTOLOGY[3][4]

Recently, ontology research is developing because of knowledge sharing and recycling knowledge. In current practice, ontology is the

term used to conceptually systematize the targeted world rigorously and exhaustively, and to create a hierarchical description of relation. Not only objects (noun), but also the process (verb), as well as constraints, are systematized.

The primary meaning of the term ontology is existence. This term is being adapted in the world of informatics to describe a target “real world” to describe algorithmically. That actual domain is a referent more powerful than a dictionary.

Moreover, ontology can identify tacit information that differs from the data of an electronic dictionary, and exists in the background of knowledge. Tacit information might alter the significance of knowledge described by the vocabulary. It is one of the important roles of ontology to clarify such tacit information.

Ontology includes general ontology that describes the concept of specializing, and domain ontology that describes the concept of specializing in the field.

4. USE CASE MODELING

UML is a typical modeling language, and 13 kinds of diagrams are defined, among them use case diagram, which is used to describe a functional side.

What kind of user (It is called actor) uses the system? Moreover, what function (It's called use case) does each actor use? This relation is described as shown in Figure 1.

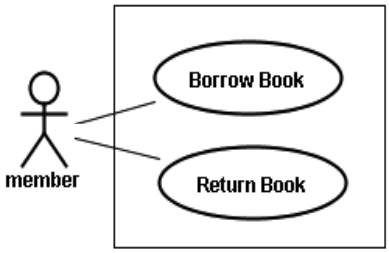


Figure 1 Example of use case diagram

In this thesis, we analyze use case diagram that students made by the information system design study. We then extracted problems in use case modeling study.

In this study, the procedure for each theme has been to lecture on knowledge necessary in each theme, studying the theme in groups, who make a presentation the next week. One of the themes is use case modeling.

In the lecture we introduce the case with a restaurant. The design of "Equipment management system of the elementary school" is given as a common challenge. The design of "Case that the group chooses" is given as a group assignment.

Figure 2 is example of use case diagrams for "Equipment management system of the elementary school" that the student made. Most groups are describing five use cases "accept reservation", "borrow equipment", "return equipment", "question & answer", and "register new equipment". These are the descriptions corresponding to the requirement, and it is extracted correctly.

However, the system doesn't work only by these functions.

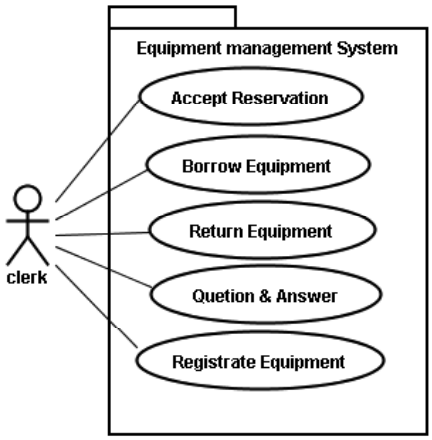


Figure 2 Example of use case diagram for equipment management system

Figure 3 is use case diagram by the teacher. Some functions not described in the requirements specification by students are added to this use case diagram. For example, it is necessary to record what to borrow, and to whom you borrow. Therefore, it is necessary to register the equipment and the user beforehand. Moreover, there is a return if there is a borrowing. A function is necessary if the item is not returned in the time limit. Therefore, the function of press is also necessary. For this the user's name, address and telephone number are required.

Moreover, it is necessary to register information for other schools (where to make contact, and what equipment can be borrowed) beforehand to make equipment available to another school. All functions to retrieve the specification in the lending situation regarding the equipment are noted.

If there is an experienced person of the information system development present, the lack

of knowledge in all the others becomes evident. This raises the problem of how to extract required functions accompanying, though not described, in the requirements specification.

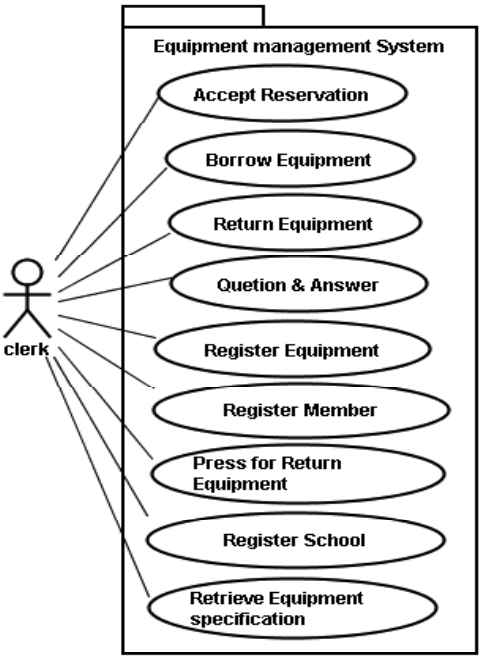


Figure 3 Model answers of use case diagram for equipment management system

5. PROPOSAL OF USECASE MODELING USING ONTOLOGY

In this proposal, the domain ontology is classified into the object ontology and the task ontology.

The object ontology defines the hierarchical relations between objects. The relation of the noun like the person and the thing, etc. is defined. The relation is described by using the "is-a" relation as shown in Figure 4.

The relation of "equivalence" is described to absorb the difference of the expression.

The task ontology is ontology that defines the attribute of the task (verb concept). It describes it with seven relational operators as shown in Figure 5.

The feature of the proposed domain ontology was to have prepared "pre-task" and "post-task" in relational operators of the task ontology. The flow of the task can be expressed by adding these relational operators. The before task and the after task are understood by referring to this ontology.

The procedure for refining the use case diagram that the student made is as follows.

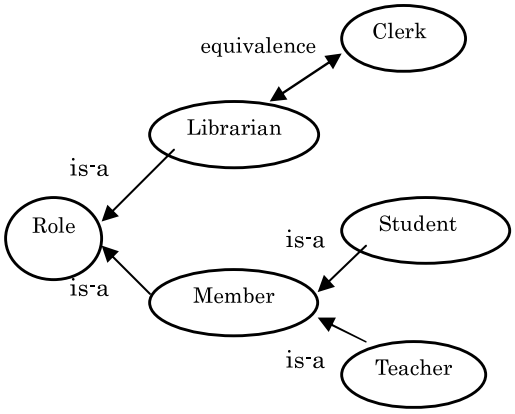


Figure 4 Composition of object ontology

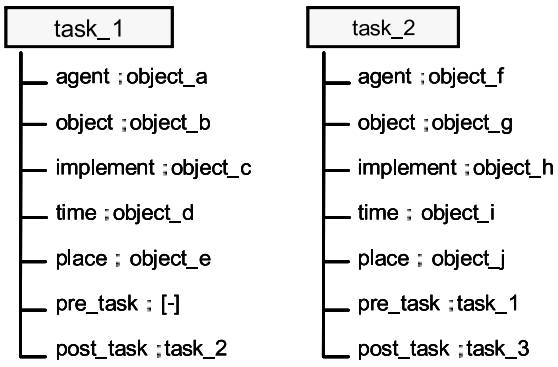


Figure 5 Composition of task ontology

At first, retrieve the task described in the use case that the student made in the beginning, and refer to the ontology. Then, the student understands what task is necessary before and behind the task.

Moreover, relational operators such as agent and object are described in each task. How the task takes part from these relational operators in the system can be judged. It is a repetitive process to trace the flow from a certain task, and to refer again to the ontology of each task. The function that the system should offer by this operation can be discovered, and the practical oversight of the use case can be discovered.

6. EXPERIMENT AND CONSIDERATION

6.1 Exercise

To verify the utility of the proposal technique, the case study of the library system was done. The requirements specification and model answers to the library information system were quoted from "Essence of the UML modeling" [6]

[Requirements specification of library information system]

There are insufficient clerical officers in charge of the lending business at the library at A-university. Then, A-university decided to introduce the library information system for the lending business efficiency.

The following three functions are necessary for the system.

- Acceptance processing of lending reservation
- Management of lending and return books
- Record of lending history

"Lending reservation" function is necessary for the user that wants to borrow the book as soon as possible.

There may be several copies of one book. Therefore, the collection of books is managed by the management indexing number, to distinguish them.

The library users are students and teachers at A-university, and all members are registered. The student and the teacher are divided because there is a restriction "Journals are not lent to the student".

Student member's lending limit is 6 books, and teacher member's lending limit is 12 books. The lending period is up to three weeks.

The object ontology and the task ontology for the library were made based on the proposal technique.

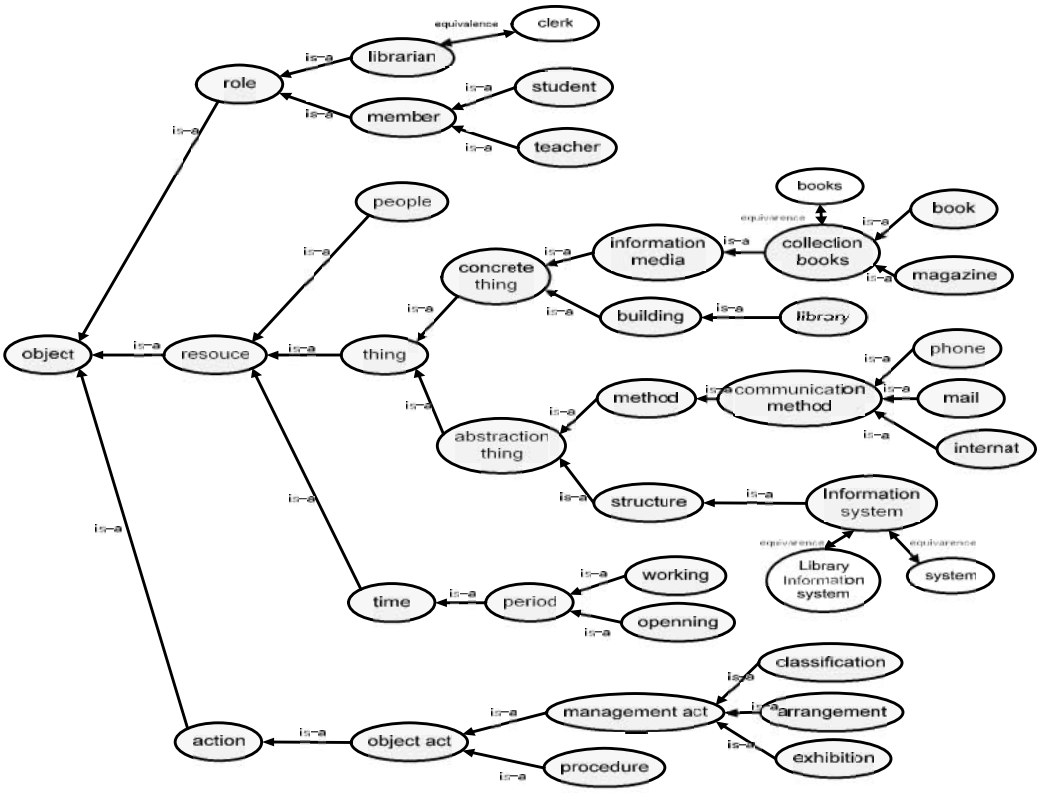


Figure 6 The library Object Ontology

The clerical officer at the library places the book in holding area when the book is returned, and reports that the book can be loaned out to the user who reserved it. It is necessary to manage which user has reserved the book to achieve this service.

The collection of books at the library is classified into books and journals. There is only one journal for each title, although there are volumes and numbers.

Figure 6 and Figure 7 show the object ontology and the task ontology. Each object and each task has extracted the verb and the noun from business manual [7] at the Shizuoka University library. And, words and phrases that were able to be used in this case were chosen.

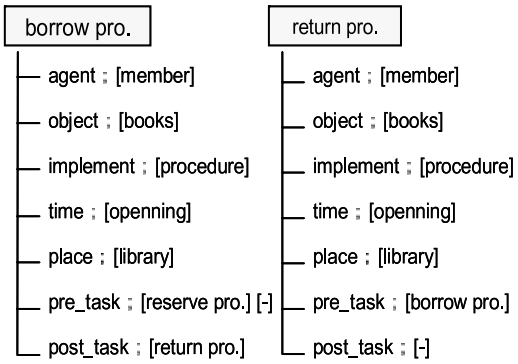


Figure 7 Example of the library task ontology

6.2 Experiment

The experiment was conducted on the following hypotheses.

[Hypothesis 1] Use case diagram made by using ontology is refined more than use case diagram made without ontology.

[Hypothesis 2] The refinement result of use case diagram made by using ontology doesn't depend on a prior lecture environment (teacher's difference, content of the lecture, numbers of students, etc).

Testees are eight beginners of UML. Four students (Hereafter, it is called group-A) have never learned UML. Other four students (Hereafter, it is called group-B) have learned some UML.

We conducted the experiment according to the following procedures.

1) We gave a lecture to the group-A concerning use case diagram.
2) The requirements specification of the library system is presented, and all testees make use case diagram (1st edition) from this requirements specification.
3) We explain the conceptual ontology.
4) The library ontology is presented.
5) Use case diagram (1st edition) is reviewed referring to the library ontology and 2nd edition is made.

6.3 Evaluation

We judged whether extracted use case was more appropriate than use case of model answers.

Hypothesis 1 was verified by the comparison use case diagrams between 1st edition and 2nd edition. Figure 8 shows the result of comparing between use case of 1st edition and use case of 2nd edition.

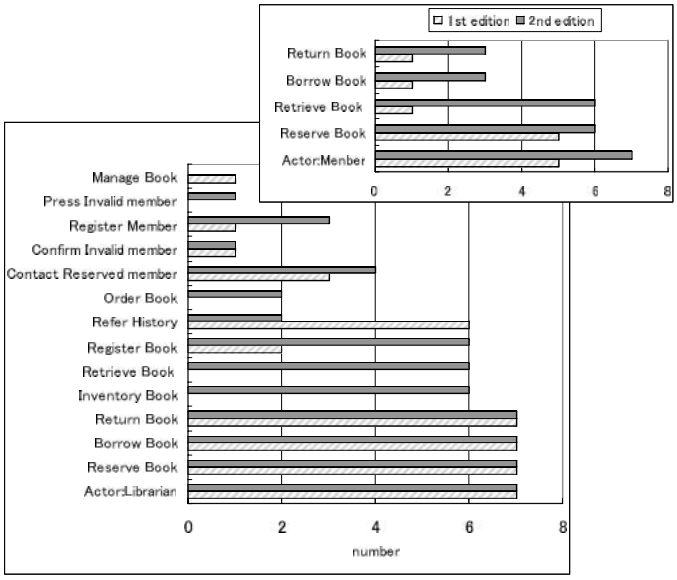


Figure 8 Comparison use case between 1st edition and 2nd edition

If the term had the same meaning even if the use case name was different, it counted assuming that it was the same. For instance, "Put back book" and "Return book" have the same meaning, and it counted as "Return book".

Use case diagram after reference to ontology contained more refined distinctions than use case diagram before reference to ontology as shown in Figure 8, and the effect of ontology was proven. However, after referring to ontology, students who extract use case "register member" are three people. This was caused by the method of describing ontology, and it is necessary to improve the description.

Next, hypothesis 2 was verified by the comparison mutually made the 1st edition and the 2nd editions by group-A and group-B.

The difference of use case in use case diagram (1st edition) is 0.5 or less in each use case as shown in Figure 9.

The difference of use case in use case diagram (2nd edition) is 0.5 or less in use case other than use case "register member" as shown in Figure 10.

It has been shown that the effect of the refinement by the ontology reference is independent of a prior lecture environment.

However, group-A has passed one year since use case modeling was studied. Group-B studied use case modeling and experimented at once. Therefore, it is necessary to verify the influence of the difference of the elapsed time after the study of modeling.

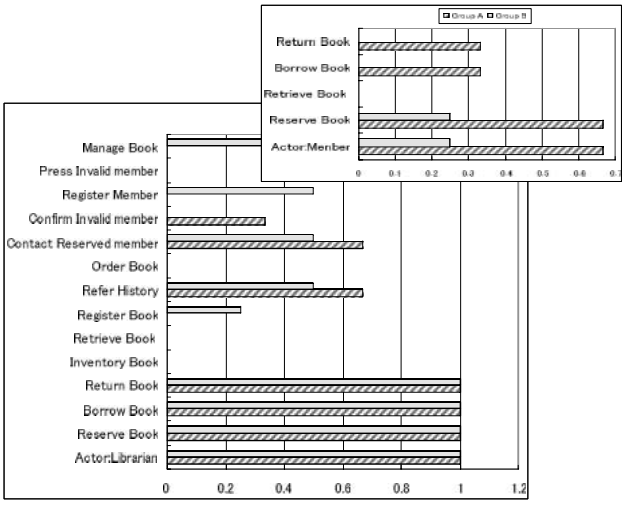


Figure 9 Comparison use case (1st edition) between group-A and group-B

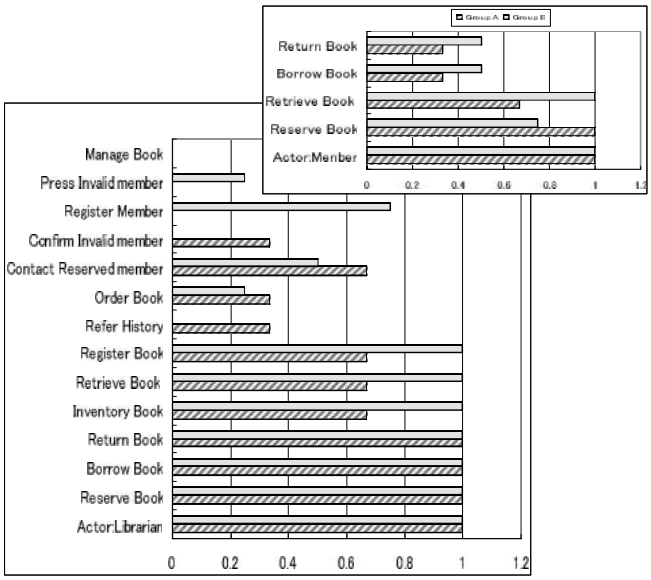


Figure 10 Comparison use case (2nd edition) between group-A and group-B

7. CONCLUSION

It has been clarified that the refinement technique based on using ontology was useful for making use case diagram.

Future tasks are as follows.

・Review of domain ontology for making more detailed refinement

・Increase the number of experiment samples, and do a statistical verification of hypothesis 1 and hypothesis 2.

・Systematize the technique for making the proposal.

The systematization of the proposal technique is a system that traces, checks the flow of ontology for making use case by building ontology into the use case diagram making tool, and finds the deficiency of use case.

The refinement level can be expected to reduce the oversight of the relating ontology by systematizing it, to be able to extract appropriate use case, and to go up.

The constructed ontology is domain ontology that specializes in the library information system. A similar system like various rental systems and reservation systems can use the ontology for enhancing effectiveness.

Future tasks are to ascertain the domain of applicability, and to add ontology smoothly to expand the domain of applicability.

REFERENCE

[1] Minegishi T., Ishibashi M., Fukuda N., Ijima T., Yamaguchi T., "Supporting Software Engineering Processes with Ontologies " , The 19th Annual Conference of the Japanese Society for Artificial Intelligence (2005).

[2] Kamiya S., Ishibashi M., Morita T., Fukuta N., Ijima T., Yamauchi T., “Supporting Analysis Class Modeling with Ontologies”, IEICE Technical Report KBSE2005-15 (2005).

[3] Ito H., “Using Ontology for Knowledge Sharing and Reuse”, UNISYS TECHNOLOGY REVIEW, No.64 (2000).

[4] Mizoguchi R. (Eds.), Furusaki K., Kimura T., Sasajima M., Mizoguchi R., “Introduction to ontology construction”, Ohm press (2006)

[5] Kodama K., “Essence of UML modeling”, Nikkei BP, (2004)

[6] NALis, operative manual, Library Hamamatsu annex attached to Shizuoka University.

A mirror-effect-based mutual tutorial
- assisting the operation learning on different interfaces for a single information service -

Keiko Katsuragawa[†‡], Takeshi Oono[†], Minoru Tomikashi[†], Satoru Kogure*,
Toshihiko Itoh**, Tatsuhiro Konishi*, and Yukihiro Itoh[‡]

[†]Nissan Motor Co., Ltd.
[‡]Graduate School of Science and Technology, Shizuoka University, Japan
* Faculty of Informatics, Shizuoka University, Japan
** Graduate School of Information Science and Technology, Hokkaido University, Japan
kei-k@mail.nissan.co.jp

Abstract - The purpose of this paper is to propose a mirror-effect-based mutual tutorial on different kind of interfaces for a single information service. When a user operates a system via a certain interface, the mirror-effect-based tutorial helps him/her learn the operating procedures of other interfaces of the same system. When conducting a certain task, the system uses the same program and input data, irrespective of which interface is used by a user. Taking advantage of this fact, our tutorial system generates operation procedures for implementing the same function on different interfaces when a user carries out a certain task. Accordingly, our tutorial involves the following two features. (1) Presenting the procedure to perform the same task from a different interface when a user operates a system via a certain interface. (2) Providing the efficient procedure depending on each interface. The task completion time is reduced by 24% and the input acceptance rate is increased by 17%.

Keywords: mirror-effect-based, mutual tutorial

1 INTRODUCTION

There are many different information services readily available in daily life today. The types of information terminals used and the environments in which users access these services are continually diversifying. In line with these trends, the interfaces for accessing information services are also becoming truly diverse.

For example, the Shinkansen ticket reservation system [1] provides two kinds of interfaces for cell phones and PCs. Many car navigation systems also have a built-in microphone for inputting voice commands while driving, in addition to ordinary control switches and a remote controller. Users can change input devices in accordance with match their circumstances.

In addition, since car navigation systems are now connectable to the Internet, Web sites, which were previously accessed mainly from PCs or mobile phones, can now also be accessed from equipment built in vehicles [2]. As a result, such information systems must now be designed to be accessible from a wide variety of terminals.

Although there have been a lot of studies on the usability of interfaces, most of those studies examine an interface accessed from a single type of device [3][4], and there have been almost no studies which examine an interface accessed from multiple types of devices.

Figure 1 shows a typical example of a system accessible from multiple interfaces. In this paper, we use the term ”device” to refer to such input equipment as a mouse or a microphone that is used with a PC, a car navigation system, etc. We further user the term ”interaction style” in the sense of Shneiderman [5]. He cites menu selection, form fill-in, command language, natural language and direct manipulation as five typical interaction styles in interactive software. Finally, we use the term ”interface” to refer to an input/output mechanism that combines the devices and the interaction styles provided by terminals.

For example, one interface combines a PC keyboard and command language, and another interface combines switches on a car navigation system and menu selection. There is also an interface that combines speech recognition through a microphone and menu selection.

Although many systems provide multiple interfaces for the purpose of selective use of multiple interfaces according to the user’s circumstances, there are virtually no examples of an interface design policy that is aimed at easy shift from one interface to another. In many cases, each interface is designed separately. Consequently, even when users operate a familiar system, they must newly learn the operating procedure if they use unfamiliar interfaces. In such a case, they must once again go through a process of repeated trial and error, just as they did the first time they operated the system. This means that we cannot improve the system’s usability by simply increasing the variety of terminals or interfaces capable of accessing a system and merely improving the individual operating ease of each device/interface.

Selective use of multiple interfaces impose a heavy burden on users, and it cannot be neglected especially when systems take advantage of a ubiquitous environment, in which people selectively use various types of information terminals according to their circumstances. For this reason, there are growing demands for interfaces such that users can smoothly shift from one to another depending on their circumstances.

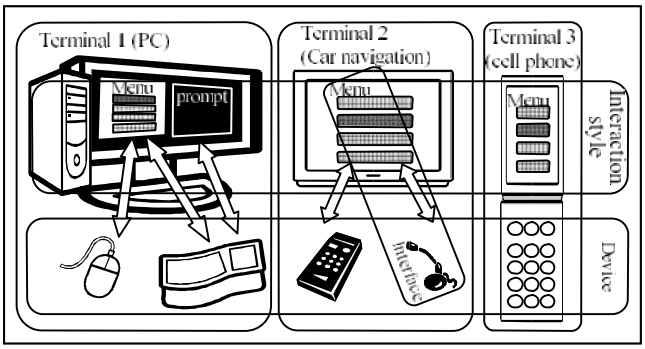


Figure 1: Example of multiple interfaces.

For in-vehicle equipment in particular, users have substantially different circumstances depending on whether they are driving or not. Unless a separate interface tailored to each of these circumstances is provided, the usability may decline markedly. In addition, for systems that are intended to be accessed both from in-vehicle equipment and from outside a vehicle (e.g. from a home PC via the Internet), it is necessary to provide an interface that is easy to use from each type of device although the circumstances are completely different from each other. Consequently, systems accessible from in-vehicle equipment in particular inevitably require some mechanism that allows users' selective use and smooth shift of their interfaces.

In this regard, this paper proposes a tutorial that helps users learn the operating procedures of unfamiliar interfaces during conducting a task through a familiar interface. When users are carrying out a certain task through a familiar interface, our system demonstrates the procedures for the same task through another interface. We call this learning support system a mirror-effect-based mutual tutorial system. By adopting this tutorial system, we can construct an operating environment for information system that fully takes advantages of multiple interfaces.

This paper provides an evaluation of the mutual tutorial function. The function is implemented onto the drive planning (DP) system constructed in our previous studies [6]. The DP system is accessible from two kinds of interfaces. One interface is designed for the natural language interaction style with a PC keyboard, and the other is designed for the menu selection interaction style with a car navigation system's switches on its instrumental panel. The usability of our system incorporating the tutorial method to be proposed is evaluated on the basis of the acceptance rate of the user input and task accomplishment time.

2 MIRROR-EFFECT-BASED MUTUAL TUTORIAL SYSTEM

We propose a mirror-effect-based mutual tutorial system as a method of improving the usability of a system with multiple interfaces. This section explains the method in detail.

2.1 Definition and Key Principles

We define the term "mirror-effect-based mutual tutorial" as follows; for every task a user actually executes via one interface, the system demonstrates how to perform the same task efficiently on another interface. In the mirror-effect-based tutorial, when a user executes a task using a certain interface, the efficient counterpart operation on another interface is automatically presented to the user as if it were a mirror image of the original operation. In this way, users can unconsciously learn how to use unfamiliar interfaces without active learning of the unfamiliar interfaces.

Figure 2 shows the basic design of the mirror-effect-based mutual tutorial system, which generates an efficient procedure on an interface which corresponds to the task executed on another interface.

The role of an interface at the time of a user input is to interpret the input, identify the necessary program and generate input data into the program. To accomplish it, each interface is equipped with an interpreter that interprets the user's input. In Figure 2, the interpreter A and the interpreter B interpret the inputs from the device A and the device B, respectively.

Generally, identification of the program set to be executed and generation of the input data based on the user's input operation are conducted uniquely. In cases where the input from each interface is aimed at executing the same task, the respective inputs are converted into the same input data and trigger the same program set inside the system. Consequently, if the system has an inverse interpreter for each interface that generates the targeted operation on the interface based on the program set to be executed and the input data for each program, the input to a certain interface feeds the system so that the system generates the corresponding operations for other interfaces. In Figure 2, the interpreter A^{-1} is the inverse interpreter that generates input operations for the interface A, and the interpreter B^{-1} is the inverse interpreter that generates input operations for the interface B.

When implementing an inverse interpreter, the following three possibilities must be considered:
(1) There is no input operation corresponding to the input data and the program to be executed.
(2) There is only one input operation corresponding to the input data and the program to be executed.
(3) There is more than one input operation corresponding to the input data and the program to be executed.

In the case (1), the task to be executed on the basis of the identified program and input data cannot be performed from the interface associated with the inverse interpreter. In this case, it is sufficient to indicate to the user that the task cannot be executed via that interface.

In the case (2), the input operation is the only output from the inverse interpreter.

In the case (3), the procedure which is expected to require the shortest operating time is selected among the operating procedures that can be generated. This enables the user to learn the most efficient operating procedure. If there is no difference in the expected operating time of the generated

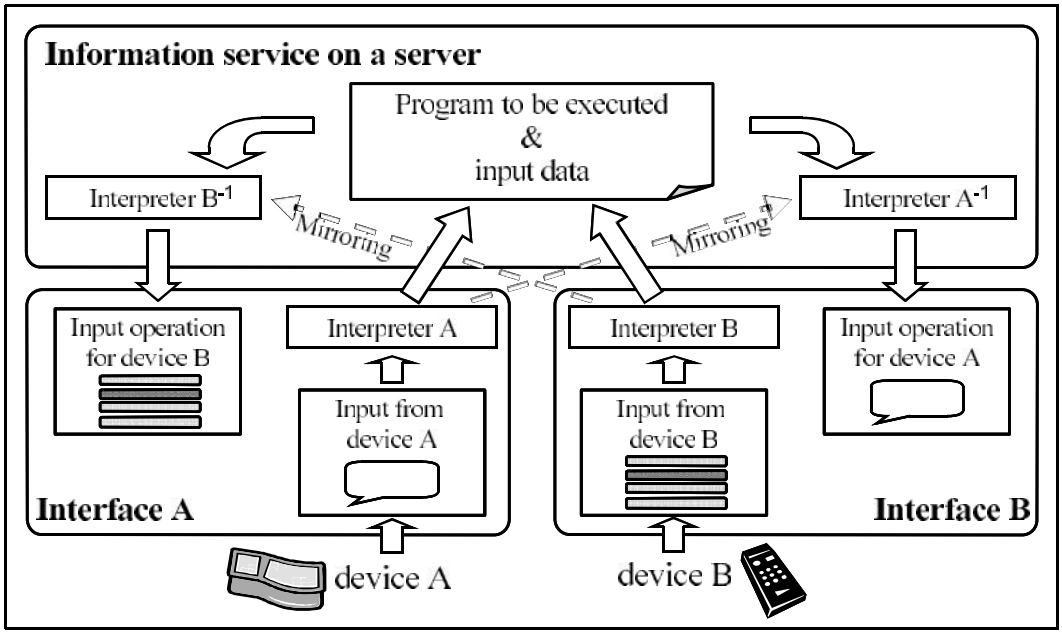


Figure 2: Basic design of mirror-effect-based mutual tutorial system.

operating procedures, a predetermined typical procedure is selected. (From the viewpoint of efficiency in this particular sense, inputting a sentence via a natural language interface is not different from inputting another sentence in which a synonymous expression is used.)

How to design inverse interpreters depends on the deference of input flexibility among interfaces. However, the number of programs to be executed is finite, and the finiteness makes it possible to define appropriate inputs for each interface to run a given program (e.g. sentence styles for natural language interface, the kind of input information and its order for menu selection interface, etc.). In addition, if the number of parameters for a given program is finite and if the specification of the parameters is definable (e.g. substituting a suitable word into an input sentence with a designated sentence style, selecting an item from a menu list, etc.), we can design an inverse interpreter for any interpreter.

2.2 Features and Expected Effects

Based on the key principles explained in the previous subsection, we can realize a mutual tutorial with the following two features.

a. When a user executes a task via a certain interface, the system teaches the procedure for the same task to be performed on a different interface.

When a user carries out a certain task, the user employs a particular strategy. In facility search, for example, there are at least two strategies: (1) specifying the genre and the landmark near the target facility and (2) specifying the genre and the address of the target facility. A user who frequently uses a particular type of strategy from one interface is expected to employ the same strategy from another interface. The mirror-effect-based mutual tutorial system demonstrates the operating procedure via another interface which corresponds to the user's strategy, only when the user actually carries out a task from one interface. This allows a user to learn operating procedures efficiently by limiting them to the functions that the person actually needs. In addition, since the tutorial is given when the user executes a task on a familiar interface, we can expect that the user learns how to use another interface while using the interface. In other words, we can expect that the user does not have to spend time to especially learn operations on unfamiliar interfaces.

b. The system teaches the most effective procedure according to each interface.

Note that operating procedures executable from different interfaces may differ in their efficiency. Some operating procedures are efficient if they are performed on a certain interface, but they are not if executed via another interface; i.e., other procedures are more efficient in the latter case. Further, there are cases where an efficient operating procedure executable from one interface is simply unavailable on another interface. These facts make it more complicated for users to learn how to operate each interface efficiently according to the task to be performed.

Taking the above considerations into account, the system teaches the most efficient operating procedure on another interface for executing the same task as the one the user actually performs. In other words, the system does not teach the operating procedure involving the same operating steps as the user's original operation if the procedure is not efficient. This feature can help users notice that the most efficient

operating procedure on one interface is not necessarily the most efficient procedure on another interface and vice versa. In Figure 2, the selection of the most efficient procedure is conducted when the inverse interpreters generate operating procedures. When an inverse interpreter can generate more than one procedure, the most efficient procedure is selected and thus the system can teach the interface-specific efficient procedure.

3 MUTUAL TUTORIAL IN THE DRIVE PLANNING SYSTEM

This section describes design examples of inverse interpreters for a menu selection interface and for a natural language interface. We also explain examples of the mirror-effect-based mutual tutorial employing these inverse interpreters.

3.1 Drive Planning System

In this study, we have selected our drive planning (DP) system [6] as an example of a system in which the same functions can be used from multiple interfaces. The mutual tutorial function has been implemented in the DP system.

The DP system allows drive plan data created on a PC to be uploaded to an online server; the system supports a variety of helpful functions including route guidance, facility search and plan editing via a car navigation system or a mobile phone. We have selected the facility search function as a target of implementing the mirror-effect-based mutual tutorial function.

When users access to the DP system from a PC, the system offers a natural language interface with keyboard input in order to take full advantage of a standard input device of a PC (i.e., keyboard) and the user- friendliness of natural language input. The natural language interface is aimed at enabling users to input their requests in whatever style they like. For example, a user can simultaneously input a complex combination of search conditions as in "Please look for a hotel near the national university in Hamamatsu".

The car navigation system interface for the DP system adopts a menu selection style manipulated by control switches on the instrumental panel. The interface also allows users to search for a facility by specifying relatively complex conditions (e.g., by designating the type of target facility and its nearby facility). Nearly all of the various facility search functions that can be used from a PC are also available on the car navigation system interface.

3.2 Generation of an Efficient Operating Procedure for Another Interface in the DP System

The facility search module of the DP system accepts search conditions like the target facility's category, address, and nearby facility. The specified search conditions are expressed in a tree structure and transferred to the search module. Figure 3 shows an example of the tree structure for a "search for a fast food restaurant near a train station in Hamamatsu City, Shizuoka Prefecture." The nodes of the tree are connected by links (solid lines in the figure) expressing upper/lower hierarchies or entire/partial hierarchies between concepts and by links (dashed lines) expressing the modifying/modified relationships between concepts. The tree nodes are expressed as frames that store pairs of an attribute and its value and/or pairs of an attribute and a pointer to its attribute value. For example, the address frame stores a pair of the attribute "prefecture" and its value "Shizuoka", and a pair of the attribute "municipality" and a pointer to its value (the address frame containing "Hamamatsu City").

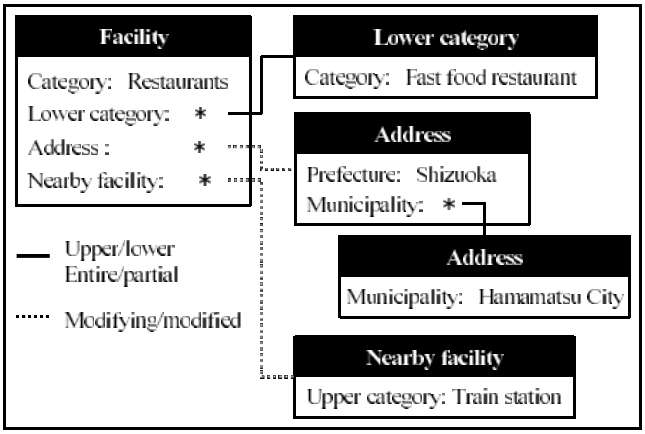


Figure 3: Example tree structure for restaurant search: "search for a fast food restaurant near a train station in Hamamatsu City, Shizuoka Prefecture".

Both interfaces use the same search module of the DP system. They transfer the same tree structure to the module if the search conditions are the same. Accordingly, the search module and the tree structure can be regarded as the source for the inverse generation of operating procedure on an interface. This is because the search module determines the program to be executed and the tree structure gives input data to the program irrespective of which interface is used.

We do not discuss the details of the conversion method from natural language input and menu selection input to the tree structure in the DP system since the discussion falls outside the scope of this paper. Conversion from natural language input to the tree structure is explained in [6]. The conversion from menu selection input to the tree structure involves defining the correspondence between each menu item and each attribute of the node in the tree, and transmitting the values selected from the menu to the corresponding attribute values in the tree structure. The following subsections explain the procedures for the inverse conversion from the tree structure to the operating procedure for each interface.

3.2.1 Inverse Generation of Natural Language Input from the Tree Structure

In generating an input natural language sentence from the corresponding tree structure, the system first extracts surface expressions corresponding to the attribute values in each frame. Facility categories and addresses are hierarchized according

to their upper-lower or whole-part relation. An expression corresponding to an upper/whole concept is connected with an expression corresponding to its lower/part concept through ”no” (the Japanese counterpart to English ”of”). If the meaning of a lower/part expression is unambiguous without the restriction imposed by its upper/whole concept, the upper/whole expression is omitted and only the lower/part expression is used so as to generate a non-redundant natural language expression.

Expressions generated from modifying concepts are then connected with the modified expressions through connecting words like ”no (of)”, ”de (in)”, ”no-chikaku-no (near)”, etc. These processes generate a natural language expression from the corresponding tree structure. In order to deal with the cases in which a certain concept is associated with multiple connecting words and is restricted by multiple modifying concepts, their linear order is predetermined according to the attributes in a given frame. The system can use different connecting words depending on whether a modifying expression immediately preceding the modified nominal, or a modifying expression is separated from the modified nominal by another intervening modifying expression.

Suppose that the tree structure in Figure 3 is generated as a result of menu selection operation. In this case, ”restaurants” and ”fast food restaurants” have an upper-lower relation and ”Shizuoka” and ”Hamamatsu City” have a whole-part relation. Since ”fast food restaurants” and ”Hamamatsu City” are unambiguous without their upper/whole concepts, the lower/part concepts are used in generation of a natural language expression. The tree structure has two modifying links to ”address” and ”nearby facility”. Their relative order is predetermined so that the address precedes the nearby facility. The connecting words used in this configuration are predetermined so that ”de” is used for the address and ”no-chikaku-no” is used for the nearby facility. As a result, ”Hamamatsu-shi-de eki-no-chikaku-no famiresu (a fast food restaurant in Hamamatsu City near a station)” is generated as an efficient input for a natural language interface.

3.2.2 Inverse Generation of Menu Manipulation Procedure from the Tree Structure

The menu selection interface is designed to put search conditions into the tree structure for facility search; that is, each choice on the menu is uniquely associated with a particular position in the tree structure. Consequently, by taking advantage of this unique relation between the menu selection and the tree structure, the system can determine which item should be selected on which menu. The system therefore can generate menu selection procedure from the corresponding tree by checking whether each attribute has its value and by determining which choice on the menu is corresponding to the attributes that have their values.

Specification of the values of a lower/partial level of hierarchized attributes can be done only if the upper/whole concepts have been specified. For that reason, the upper/whole concepts are specified first. The order for specifying other attribute values is basically arbitrary, but it is more efficient to specify them in the order in which they are displayed on the screen. In cases where the same value can be specified for multiple selection operations, the procedure with the fewest operations is selected.

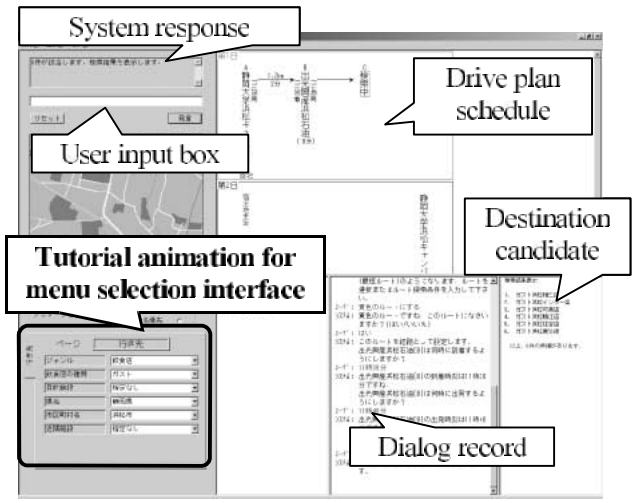


Figure 4: Example tutorial of the menu selection interface during operation on the natural language interface.

For example, the system generates menu selection procedure from the tree structure in Figure 3 as follows. Since ”Shizuoka” and ”Hamamatsu City” have a whole-part relation, ”Shizuoka” is selected on the prefecture menu before the selection of ”Hamamatsu City” on the city menu. Since ”restaurants” and ”fast food restaurants” have an upper-lower relation, ”restaurants” is first selected on the genre menu. Then the restaurant category menu becomes available and ”fast food restaurants” is selected on the restaurant category menu. The facility category, the address, the nearby facility (station) can be specified in any order. However, since they are displayed on the screen in the order of the facility category, the address, the nearby facility, the system generates the menu selection procedure with this particular order.

3.3 Mutual Tutorial in the DP System

As the tutorial of the menu selection interface while operating on the natural language interface of the PC, we have adopted an explicit method whereby the actual input procedure is shown in an animation. The PC system has a menu display for tutorial that is identical to that of the car navigation system. Suppose that a user enters ”Hamamatsu-shi-no famiresu-ni iki-tai (I want to go to a fast food restaurant in Hamamatsu City)” in the PC system. The menu display on the PC then shows an animation that shows how to conduct facility search with the same search conditions on the menu selection interface. The animation enables the user to directly watch which choices should be selected on which menu in what order. Figure 4 shows an example of a tutorial of the menu selection interface during operation on the natural language interface.

The tutorial of the natural language interface during op-

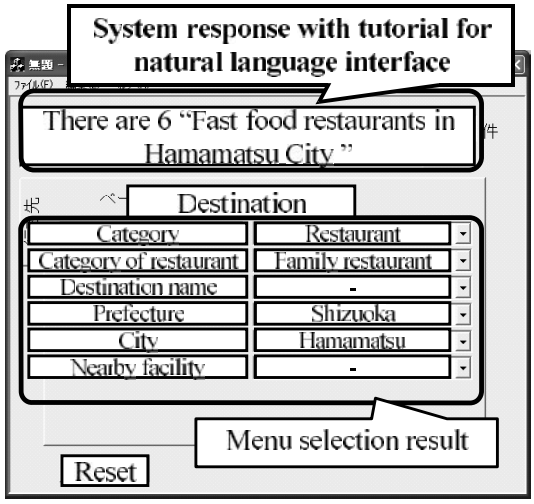


Figure 5: Example tutorial of the natural language interface during operation on the menu selection interface.

eration on the menu selection interface adopts the following method. When a user conducts facility search on the menu selection interface, the DP system gives a response sentence to confirm the search conditions. In generating the response sentence, the system use natural language expressions suitable for the most efficient search on the natural language interface and those expressions are highlighted in the response sentence. Suppose that a user selects "restaurants", "fast food restaurants", "Shizuoka" and "Hamamtsu City" on the genre menu, the restaurant category menu, the prefecture menu and the city menu, respectively. The system then replies "Hamamatsu-shi desu-ne. **Hamamatsu-shi-no famiresu-**wa 24-ken desu. (You have selected Hamamatsu City. There are 24 **fast food restaurants in Hamamatsu City**.)". In this reply, "Hamamatsu-shi-no famiresu (fast food restaurants in Hamamatsu City)" is highlighted. In repeating such tutorials, users unconsciously acquire natural language expressions suitable for the most efficient search on the natural language interface. Figure 5 shows an example of a tutorial of the natural language interface while operating on the menu selection interface.

When a user uses a function that is unavailable on another interface, the system must show the user that the function is unavailable on another interface. In our DP system, detailed editing of a drive plan such as setting the departure/arrival time can only be done from the PC system, and these operations are not available on the car navigation system's menu selection interface. When a user uses such functions on the PC system, the tutorial system shows their unavailability by not showing an animation of the operations on the menu selection interface.

4 EVALUATION OF THE MUTUAL TUTORIAL SYSTEM IN THE DP SYSTEM

This section describes the evaluation of the effectiveness of the mutual tutorial system.

4.1 Effect of the Tutorial of the Natural Language Interface During the Operation of the Menu Selection Interface

First, a car navigation system without a response-based tutorial function was selected as the system for comparison. The mutual tutorial function was incorporated in another car navigation system, and the effect of the tutorial on the operation of the PC system interface was investigated. We have evaluated the effect of the tutorial of the natural language interface during the operation of the menu selection interface. The operation results of the two user groups have been examined. One group operates the natural language interface after using the car navigation system with the tutorial function, and the other group operates it after using the car navigation system without the tutorial function.

4.1.1 Experimental Procedure

The subjects were twenty engineering students and they were divided into two groups. Group NT first used the car navigation system without tutorial function and then used the DP system on a PC. Group T first used the car navigation system with the tutorial function and then operated the DP system on a PC. The two car navigation systems have no difference except the tutorial function. In order to familiarize the subjects with their respective car navigation system, we made them practice operating the system for 30 minutes every day over a three-day period. The subjects then subjectively evaluated the car navigation system that they used. After that, a prepared conversation of a certain family talking about their desires regarding a family trip was presented to the subjects. All of the subjects then used the same PC system to create a drive plan that satisfied the desires of the family members.

The conditions under which searches could be conducted, and the operations particular to the DP system (e.g., halt a search midway by entering the command "stop search") were explained to the subjects beforehand using the same manual sheet alone.

4.1.2 Experimental Results

After creating the drive plan, the subjects evaluated the PC system subjectively by responding to a questionnaire using a 7-point evaluation scale. An objective evaluation was also made on the basis of the input acceptance rate, the number of input times, the number of key words per sentence and the task accomplishment time.

As the first step, a questionnaire was conducted among the members of both groups concerning their respective car navigation systems. The purpose of the questionnaire was to confirm that there were no differences between the two navigation systems in terms of their performance.

The questionnaire results of the two groups and the results of a t-test between the two groups are shown in Table 1. The mean scores for Question 1 concerning the ease of use (easy to use = 7) are 4.9 for Group NT and 5.3 for Group T. The mean scores for Question 2 concerning the kindness of the responses (kind = 7) are 4.9 for Group NT and 5.2 for Group

Table 1: Questionnaire results for the two car navigation systems.

Questions	NT	T	p-value
1.Ease of use of system	4.9	5.3	0.419
2.Kindness of responses	4.9	5.2	0.624
3.Uncertainty about what to input	3.6	3.4	0.633
4.Ease of understanding of screen display	5.0	5.6	0.382
5.Discomfort from the response sentences	—	4.5	—

T. Those for Question 4 about the ease of understanding the screen display (easy to understand = 7) are 5.0 for Group NT and 5.6 for Group T. Although the scores for Group T are slightly higher, the t-test indicates that there is virtually no significant difference between the two groups. The evaluation scores for Question 3 concerning the uncertainty about what to input (no uncertainty = 7) are nearly equal to each other. Thus the questionnaire results indicates that there is no difference between the car navigation systems in terms of the impression the subjects received when using their respective systems.

Table 2 shows the questionnaire results after the subjects made the drive plan on the PC system following the use of their respective car navigation systems. The scores for the questions 1, 2 and 3 show significant difference between the two groups. The mean scores of Group NT and Group T are 3.9 and 5.4 for Question 1, and 4.1 and 5.3 for question 2, respectively, indicating that higher evaluation scores have been given by Group T that used the car navigation system with the tutorial function. Presumably, the effect of the tutorial from the car navigation system makes it clear what sentences should be entered via the natural language interface when searching for a certain facility. If this is the case, it surely improves the overall ease of use of the system, thereby accounting for Group T's higher evaluation scores.

The mean score of Group T for Question 3 concerning the uncertainty about what to input (no uncertainty = 7) is also higher at 3.8 than 2.3 for group NT. The fact that Group T gave a higher evaluation to the PC system also indicates the effect of the tutorial on this group. Group T also gave a high mean score of 6.1 in response to Question 5 about whether the navigation system responses were helpful (helpful = 7). This question was only given to Group T. The high score indicates that the subjects actually felt the effect of the tutorial.

Table 3 shows the results of the objective evaluation. Significant difference is observed between the two groups for all of the evaluation items. Item 1 in Table 3 shows the total time required to create the drive plan. The other evaluation results only pertain to the input process for conducting a facility search and the other functions available on the PC system alone (e.g., a route search function, a function for setting the desired arrival time, etc.) are excluded from the evaluation. This is because the facility search function is the only function that is supported by the tutorial function.

Table 2: Questionnaire results for the PC system.

Questions	NT	T	p-value
1.Ease of use of system	3.9	5.4	0.044
2.Kindness of responses	4.1	5.3	0.046
3.Uncertainty about what to input	2.3	3.8	0.002
4.Ease of understanding of screen display	5.0	5.5	0.486
5.Helpfulness of the car navigation system response	—	6.2	—

Table 3: Objective evaluation results for the PC system.

Evaluation items	NT	T	p-value
1.Plan creation time(min:sec)	13:23	10:09	0.016
2.Number of input times	24.3	11.2	0.005
3.Number of key words per sentence	1.8	2.6	0.002
4.Acceptance rate of input sentences(%)	72.4	89.7	0.005

The mean time spent by Group T to create the drive plan is 10 minutes 9 seconds, which is more than 30% shorter than that of Group NT (13 minutes 23 seconds). Item 2 is the mean number of data inputs for conducting a facility search. The result for Group T is 11.2 times, which is less than half the number for Group NT (24.3). The results indicate the efficiency of the tutorial. The efficiency is explained in terms of Item 3, where Group T put more key words into an input sentence than Group NT whereby Group T needed fewer sentences than Group NT in making the facility search. Key words here are those words that express search conditions like facility categories, prefectures/cities of the destinations.

Item 4 represents the rate at which the system correctly understood the intention of the user inputs. Group T had a better result at 89.7% compared with 72.4% for Group NT. A good reason for this difference is that Group NT used many facility search conditions that the current system does not accept, whereas Group T almost never used such inappropriate conditions. The system does not accept search conditions that are not stored in the data base (e.g., yasui hoteru (cheap hotel)), conditions that involve the purposes of search alone (e.g., shokuji-suru (have a meal)), and so on. The difference between the two groups can be attributed to the effect of the tutorial from the car navigation system; that is, the tutorial makes it clear what search conditions can be used.

The results indicate that the tutorial greatly helps users unsure about what to input on the natural language interface, by putting efficient natural language inputs for the PC system into the responses from the car navigation system.

Table 4: Questionnaire results for two PC systems.

Questions	NT	T	p-value
1.Ease of use of system	5.4	4.9	0.358
2.Kindness of responses	4.5	5.1	0.399
3.Uncertainty about what to input	4.6	4.1	0.221
4.Ease of understanding of screen display	4.8	5.3	0.540
5.Discomfort from the displays animation	—	3.9	—

Table 5: Questionnaire results for the car navigation system.

Questions	NT	T	p-value
1.Ease of use of system	4.7	5.1	0.587
2.Kindness of responses	5.0	4.6	0.529
3.Uncertainty about what to input	3.3	2.9	0.540
4.Ease of understanding of screen display	5.1	5.1	—
5.Helpfulness of the displayed animation	—	5.9	—

4.2 Effect of the Tutorial of the Menu Selection Interface During the Operation of the Natural Language Interface

We also investigated the effect of showing an animation on the PC system screen as a tutorial for the operation of the car navigation system.

4.2.1 Experimental Procedure

The test subjects were twenty engineering students who were divided into two groups, as was done in the experiment described in the preceding section. Group NT first used a PC system without any tutorial function before using the car navigation system. Group T first used a PC system incorporating the tutorial function before using the car navigation system. Before the experiment, each group used its respective PC system for 30 minutes every day over a three-day period and then answered a questionnaire. After that, all of the subjects were asked to read the manual of the car navigation system. They were then asked to execute six types of facility search tasks using the same car navigation system. One task, for example, was to ”search for a hospital near the present location.” The subjects then evaluated the car navigation system subjectively by responding to a questionnaire, and an objective evaluation was made based on their facility search time.

The test subjects were twenty engineering students, and they were divided into two groups as in the experiment in the preceding section. Group NT first used the PC system without the tutorial function before using the car navigation system. Group T first used the PC system with the tutorial function before using the car navigation system. Before the experiment, each group used its respective PC system for 30 minutes every day over a three-day period and then answered a questionnaire. After that, all of the subjects were asked to read the manual of the car navigation system. They were then asked to execute six types of facility search tasks using the same car navigation system. One task, for example, was to search for a hospital near the present location. The subjects then evaluated their respective PC systems subjectively by responding to a questionnaire, and an objective evaluation was made based on their facility search time.

4.2.2 Experimental Results

The subjective evaluation was done by a questionnaire using a seven-point evaluation scale after the subjects had completed the 6 facility search tasks. The objective evaluation was based on the task accomplishment time.

First, a questionnaire was conducted to confirm that there was no difference between the two PC systems in terms of their performance.

The questionnaire results and the results of a t-test between the two groups are shown in Table 4. Group T was asked whether they felt any discomfort from the displayed animation (no discomfort = 7). The mean evaluation score of 3.9 was slightly worse than the median. However, the scores for the other questions were virtually identical. This indicates that the tutorial animation did not influence the subjective evaluation of the PC systems.

Table 5 shows the subjective evaluation results for the car navigation system. No large difference was seen in the questionnaire results between the two groups. However, the mean score of Question 5 for Group T is 5.9. The question asked if the animation displayed on the PC system screen was helpful in operating the car navigation system. This indicates that Group T was actually aware of the effect of the tutorial.

The time needed to complete each facility search task is shown in Table 6. Since Task 1 is very simple to execute, no large difference was seen between the two groups. Task 2 is slightly more complicated because it requires the specification of a reference location. The difference between the two groups tends to increase as the complexity of the task becomes large. Task 3 requires approximately twice as many operations as the tasks 1 and 2 because multiple operations are needed to specify the reference location. A significant difference is seen between the two groups with respect to this task. Task 4 involves nearly the same operating procedure as Task 2. Since this task was performed soon after the subjects experienced Task 2, it is presumed that the effect of being familiar with the procedure resulted in there being no difference between the two groups. Task 5 is even more difficult to execute as it requires specifying the travel time (or distance). A significant difference is also seen between the two groups with respect to this task. Task 6 involves more complicated specification of the reference location than Task 5. A difference is seen in the mean task accomplishment time, although it is not statistically significant.

These results show that the tutorial has the effect of reducing the difficulty of learning the menu selection operating procedure, especially for complicated operations.

Table 6: Time required for facility search using the car navigation system.

Destination	NT	T	p-value
1.Gas station in Hamamatsu City	0:47	0:37	0.529
2.Hospital near the present location	0:40	0:30	0.142
3.Hotel near Hamanako Paruparu amusement park	2:08	1:03	0.028
4.Park near the train station	1:10	1:12	0.905
5.Lawson convenience store within 30min. from the present location	1:14	0:47	0.026
6.Police station within 10 km from Hamamatsu Castle	2:13	1:35	0.256

5 CONCLUSION

This paper has proposed a mirror-effect-based mutual tutorial system as a solution to the problem of having to separately learn the specific procedure for using a different interface every time a user changes an input device. This problem emerges when users use a system that can be operated from multiple information terminals with different interfaces. Even though a user has learned how to operate such a system from a certain interface, the user may well be unfamiliar to the procedure for operating the system from a different interface, and have to learn the operating procedure anew. Experiments were conducted to evaluate the mutual tutorial function incorporated in a DP system. The results obtained have shown that the mutual tutorial system is effective in shortening the task accomplishment time and in improving the input acceptance rate, among others.

The effectiveness of the mirror-effect-based mutual tutorial depends on how easily each interface is understood. If an interface is very easy to understand, the users would easily learn how to use it without the mutual tutorial. On the other hand, if an interface is very complicated and hard to understand, the users might not acquire how to use it through the mutual tutorial alone. Although we have confirmed the effectiveness of the mirror-effect-based mutual tutorial in the DP system, we may need further experiments using interfaces with various degrees of complexity (difficulty) in order to clarify the effective range of the proposed mirror-effect-based mutual tutorial.

The proposed mutual tutorial system adopts a framework that takes advantage of operating the same system; a user may well use the same set of functions irrespective of the interfaces he/she actually uses. By enabling users to efficiently learn operations of preferred functions on multiple interfaces, they are expected to be aware of the merits of using the same system from different interfaces. As a result, the same system is expected to be chosen by users irrespective of terminal types.

As a ubiquitous environment is steadily put in place, it is predicted that there will be increasing opportunities for using the same system from a variety of devices depending on the user's circumstances. These devices will include mobile terminals such as cellular phones and personal digital assistants (PDAs) as well as PCs. In such an environment, there will be an even greater necessity to be able to use multiple interfaces selectively and smoothly. The mirror-effect-based mutual tutorial system proposed in this paper should be effective toward that end.

REFERENCES

[1] Express E-reservations, http://expy.jp/top.php
[2] Nissan CARWINGS, Car Navigation Guide, http://drive.nissan-carwings.com/WEB/
[3] J. Nielsen, Usability Engineering, Academic Press (1994).
[4] D. A. Norman, The Psychology of Everyday Things, Basic Books (1988).
[5] B. Shneiderman, Designing the User Interface: Strategies for Effective Human-Computer Interaction, Addison-Wesley (1987).
[6] K. Katsuragawa, et al., Construction and Evaluation of a Sub System DPS-PC Which Supports Users in Making and Editing a Drive Plan, Transactions of the Information Processing Society of Japan, Vol. 44, No. 12, pp. 2990–3001 (2003) (in Japanese).

Session 4: Design Methodology (Chair Jun SAWAMOTO)

A Method to Generate Pre- and Postconditions for Java Methods Based on Hybrid Analysis

Kozo Okano†, Shinji Kusumoto†, and Yasunobu Kajita†

†Graduate School of Information Science and Technology, Osaka University, Japan
{okano, kusumoto, y-kajita}@ist.osaka-u.ac.jp

Abstract - Daikon automatically generates assertions such as preconditions, postconditions, and invariants from Java programs. Daikon executes Java programs based on test cases provided by the user, and outputs all the properties as assertions, which are matched to the resulting trace values of the variables. However, Daikon generates false assertions if the test cases are insufficient. Since it is difficult for the user to provide sufficient test cases, Daikon will inevitably generate false assertions. Thus, this paper proposes a method that generates reliable assertions resulting from insufficient test cases. The basic idea is to apply Daikon only to loop sections, where it is difficult to generate assertions based on static analysis. For the other sections of the program, static analysis is applied. Applying this method to real code shows that the method generated far fewer false assertions than Daikon alone.

Keywords: Design by Contract (DbC), Java, Precondition, Postcondition, Daikon, JML

1 Introduction

The objective of Design by Contract (DbC)[1] is to improve re-usability, quality, and reliability of software by considering software specification of class clients and class suppliers as contracts.

Checking that a program satisfies its contract occasionally uses boolean expressions called *assertions*. Assertions are associated with a method or a class and consist of several different types. *Preconditions* for a method are assertions for the arguments of the method and instance variables that should be enforced before the execution of a given method. *Postconditions* for a method are assertions that should be enforced after the execution of the method. Finally, *class invariants* are assertions that the given object should observe during its life. Such assertions abstract the behaviour of a method and an object of the class. Thus, assertions can also increase program comprehension due to their formal description style.

Eiffel[2] is an example of a programming language that implements DbC. As well, DbC can be found in Java. For example, J2SE 1.4 introduces the assert sentences in its syntax and provides assertion. The Java Modelling Language (JML)[3], [4] also proposed a formal specification language for Java.

JML allows assertions to be written in the comment area of the Java source code. Such assertions have Java-like syntax. Several tools[9] are available for JML, such as a runtime checker[5], a static checker[6], a formal verifier[7], and a model checker[8].

On the other hand, automatic generation of formal specifications from Java programs in order to reduce users' workload[10], [11] have been proposed. Daikon[13], [14] is one such implementation that automatically generates assertions based on the source code.

Daikon first identifies the *program points* of the target program. Using test cases provided by a user, it executes the target program and obtains a log file including the trace of the program variables' values at the program points. Next, it compares each trace with assertion patterns that are prepared as assertion templates. Finally, it outputs the matched assertion patterns as assertions at each of the program points.

Daikon can handle many different target programming languages, including Java.

However, this approach has the drawback that the quality of the assertion depends on that of the test cases. Since the output assertions occasionally form complex expressions, this causes readability to decrease. For example, assume that the assertion $x > 0$ is desired, but that the test cases contain only even values of x, then Daikon might generate a useless assertion like $x > 0$ && even(x).

In order to reduce the output quality dependence on the test cases for Daikon, this paper proposes a new method where Daikon is applied to limited sections of the target program, and *static analysis* is applied to the other sections of the program. Thus, Daikon considers only the loop sections of the target program, for which static analysis is relatively difficult to apply. The remaining sections are statically analysed. Typical exceptions are also analysed in the static analysis phase. Negations of such conditions are reported as preconditions. Pre- and postconditions generated using this method are more useful than use of Daikon alone.

The paper is organised as follows. Section 2 provides the related works, while Section 3 gives the details of the proposed method. Section 4 shows the experimental results. Finally, Section 5 concludes the paper.

2 Related Work

Paper [11] provides semi-automatic generation of assertions using machine learning techniques, while the proposed method uses a hybrid approach. In [10], the researchers proposed a tool called Houdini, which automatically generates assertions based on full static analysis using ESC/Java.

ESC/Java and its successor ESC/Java2 [6] can statically check that a Java program satisfies the annotated JML specification. It performs model checking for every method. It also

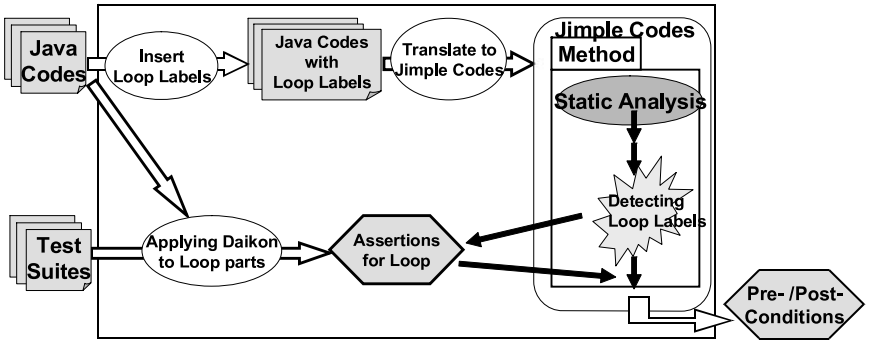


Figure 1: Tool Process

produces counterexamples when the program code does not satisfy the annotated JML specification.

Paper [12] provides fully symbolic execution analysis of a Java program (Java Card). For the loop section, it uses unwinding inference and semi-automatic weak condition generation.

In this paper, Daikon is used as the tool for the analysis of loop sections due to its lightweightness and scalability.

3 Proposed Method

3.1 Outline of the Method

In the proposed method, Daikon works only on the loop sections of the target program. For the remaining sections, static analysis (*symbolic execution*) is performed. Symbolic execution traces the incremental changes of variables' values as the commands are carried out.

At the end point of a method, the trace results are transformed into a logical expression. The resulting expression is a postcondition for the method. In general, such symbolic execution cannot be complete. Thus, only simple source code can be considered.

If a condition that causes some major exception is found, the negation of this condition is treated as a precondition for the method.

3.1.1 The Program's Class Restriction

The source code classes are restricted as follows:

- The method can only deal with basic types.
- Recursive calls cannot be analyzed.
- Cast and dynamic binding are not supported.

Reference variables are not fully analyzed but partially analyzed. This is described in detail later.

3.1.2 Overview of the Tool's Flow

Figure 1 shows an overview of the tool that implements the proposed method.

The tool performs the following processes (Preparation) before analysis:

1. It transforms the input Java program into the Jimple form using Soot[17].
2. It then applies Daikon to each of the program's loop sections and stores the generated assertions for each of the sections.
3. It inserts loop labels into the loop sections of the Jimple code in order to identify the sections.

After having performed the Preparation stage, the tool statically analyzes each of the methods in the Jimple code, and generates pre- and postconditions for the corresponding methods. In the analysis phase, the assertions generated during the Preparation stage by Daikon are analysed.

3.2 Preparation Stage

This section provides a detailed explanation of the Preparation stage.

3.2.1 Transformation into the Jimple Format

The tool does not perform the analysis using the original Java source code, but using the Jimple formatted code files. The transformation is performed using Soot[17]. Soot is a framework for Java code optimization that provides the environment and intermediate files by which the original Java source code can be transformed in the Jimple code. Jimple is one such intermediate file in which commands are represented using the 3-address format with types. This format helps to retrieve the value of target variables at any state, as well as to extract the control flow of the target method.

3.2.2 Applying Daikon to Loop Sections

As Daikon cannot use the Jimple format, the tool applies Daikon to the original Java source code. Daikon's objective is mainly to generate pre- and postconditions, as well as invariants. Therefore, it limits its program points where assertions are associated. In particular, it is desired to generate assertions at loop sections, which Daikon does not regard as program points. In order to apply Daikon to loop sections, the tool transforms each of the loop sections into an equivalent method, which will be referred to as a proxy method.

In order to preserve the equivalence between a loop section and its proxy method, a proxy variable is introduced for each of the variables in the loop section, except for variables declared in the section itself. Such proxy variables are declared in the class as field variables. Before calling the proxy method, the values of such variables are assigned to the corresponding proxy variables, and, after calling the proxy method, the values of proxy variables are assigned to the original variables.

If return statements exist in a loop, an exception is generated in the proxy method. The exception is caught at the caller method, and it executes the return.

3.2.3 Insertion of Loop Labels

The method has to identify loop sections in the Jimple code. However, in the Jimple code, a loop section is represented as a combination of labels and goto statements. Since, the tool has to refer to assertions from the loop sections, the tool must record the relationship between the assertions of each proxy method, the location of the corresponding method, and the location of corresponding loop sections in the Jimple code.

In order to resolve these problems, the tool inserts a loop label containing the ID of the proxy method, before and after the loop section in the original Java code. Such preparation helps identify the loop in the Jimple code, as well as provide a reference between the loop sections and the proxy methods.

3.3 Generating Postconditions

3.3.1 Symbolic Execution

For the target methods, the proposed process traces every local variable and its field variables symbolically. This process is called symbolic execution. The resulting relations form the conditions. The execution of a single command from the Jimple code is considered to be a state. Symbolic execution starts at state 0, which is the start point of the target method, and sequentially performs the process. The process deals with only basic types of Java programs. However, in the future, it is expected that this method will be expanded to deal with reference variables.

In this paper, symbolic execution will be illustrated with the following example. Figure 2 shows the example. Figure 3 is the Jimple code derived from Fig. 2 using Soot. "<number>" in Fig.3 shows a state. Figure 3 has 8 states (state0 to state7). It should be noted that Soot does not generate the state numbers for the Jimple code. The authors added the state numbers in order to explain the process.

```
public int test(int x, int y){
    int ret;

    if(x > y){
        ret = x - y;
    }else{
        ret = y - x;
    }
    return ret;
}
```

Figure 2: A Java method

```
public int test(int, int)  {
        IfTest this;
        int x, y, ret;

<0>   this := @this: IfTest;
<1>   x := @parameter0: int;
<2>   y := @parameter1: int;
<3>   if x <= y goto label0;
<4>   ret = x - y;
<5>   goto label1;
label0:
<6>   ret = y - x;
label1:
<7>   return ret;
}
```

Figure 3: The method translated into Jimple from the Java source code given in Fig. 2

The target variables are x, y, and ret. The variable `this` shows the instance itself. Thus, it can be ignored. Similarly, state 0 can be ignored. Starting from state 1, first the value of the first argument is assigned to x. In this case, the variable of the first argument is x. Thus, the expression $x == \backslash\text{old}(x)$ is obtained. Here, \old() stands for the value just before the method is called. In state 2, the expression $y == \backslash\text{old}(y)$ is obtained. In state 3, case-split is performed due to the if statement, and two conditions are performed. The premise clause of the first and last conditions will be the condition of the if statement. The premise clause of the former is the negation of the "if" condition, as it shows the condition of the "else" clause. The consequence clause of the first and last conditions will be generated during state 6 and states 4 and 5.

The condition of the if statement is $(x \leq y)$, and the values of x and y are $\backslash\text{old}(x)$ and $\backslash\text{old}(y)$. Thus, the condition becomes $(\backslash\text{old}(x) \leq \backslash\text{old}(y))$. In state 4, the assignment of ret is performed. In state 4, neither the value of x nor y is changed,

Table 1: Assertion Expression Syntax

assertion	::=	expr | condition "==>" expr
expr	:=	orig_expr | const_expr
condition	::=	orig_var op orig_var
		| orig_var op const
		| const op orig_var
		| const op const
orig_expr	::=	var op orig_var
const_expr	::=	var op const
orig_var	::=	"\orig(" var ")"| "\old(" var ")"
var	::=	proxy variable
const	::=	constant
op	::=	"==" | ">" | "<" | ">=" | "<=" | "! ="

(\orig() has the same semantics as JML's \old().)

and, thus, the values of x and y are \old(x) and \old(y), respectively. It can be concluded that the value of ret is \old(x) $-$\old(y).

As was stated above, the negation is added to the condition to give the condition $\neg(\backslash\text{old}(x) \leq \backslash\text{old}(y)) \Longrightarrow \text{ret} == \backslash\text{old}(x) - \backslash\text{old}(y)$, where "$\Longrightarrow$" stands for imply.

State 5 can be ignored. During state 6, the following condition is obtained $(\backslash\text{old}(x) \leq \backslash\text{old}(y)) \Longrightarrow \text{ret} == \backslash\text{old}(y) - \backslash\text{old}(x)$. In state 7, it is found that the value of rerur is ret. Thus, the condition for ret will be the consequence.

Finally, the following conditions are obtained as the postconditions for the target method:

- $\neg(\backslash\text{old}(x) \leq \backslash\text{old}(y)) \Longrightarrow \backslash\text{result} == \backslash\text{old}(x) - \backslash\text{old}(y)$
- $(\backslash\text{old}(x) \leq \backslash\text{old}(y)) \Longrightarrow \backslash\text{result} == \backslash\text{old}(y) - \backslash\text{old}(x)$

, where \result shows the return value.

Table 1 shows the expression class.

The assertions generated by the tool, along with the assertions coming from Daikon, are given in Table 1.

A simple assertion is an assertion with no condition in 1, while an equality assertion is an assertion with the equal relationship ==.

3.3.2 Assignments

For the assignment statement $v = e$, assume that the value of v has already been obtained.

If the condition for v' is a simple equality assertion $v' = e'$, then the new condition is $v == e[e'/v']$. If the condition of v is not a simple, equality assertion, $C \Longrightarrow v == e[e'/v']$ is obtained, where $C \Longrightarrow v' == e'$ is a relation of v'.

Static analysis uses only equality assertions. Thus, the above process is closed. However, if there are loop sections, such inference might not close. In such cases, the inference is aborted. Inference is also aborted for other cases, such as when variables are not basic types.

3.3.3 If and Switch Statements

In Jimple code, an if statement and a switch statement in Java are presented as a combination of goto statements. During symbolic execution, when a goto statement is encountered, conditional expressions are generated. During a state where a condition CP has already been obtained, and the condition of the goto statement is C, then the resulting condition $CP \wedge C$ is produced for the state where goto specifies. The next state of goto (the state reachable when the condition is false) has a condition $CP \wedge \neg C$. Such conditions locate the premise clause of the target state's conditions.

At the confluent state, such premises sometimes cancel each other, for example, premises C and $\neg C$, and the consequences are the same. In this case, the consequence is only used as a result.

3.3.4 Loop Sections

For a loop section, Daikon is used. The loop sections are identified with the loop labels inserted during the Preparation stage. The class of generated assertions by Daikon are restricted to those given in Table 1.

3.3.5 Method Calls

This section describes the process for other method calls.

In general, after calling a method, the postconditions of the called method and other partially evaluated conditions for the variables in the target method are joined. However, performing this combination can occasionally produce complex conditions, which decrease readability.

Therefore, additional variables are created, and the intermediate results are stored in them.

This process is explained using a simple example. A methodCall in Fig. 4 calls a method test in Fig.2. These methods are in the same class.

```
public int methodCall(int i){
  int j = 1;
  int ret;

  ret = this.test(i, j);
  return(ret);
}
```

Figure 4: A method calling the test method in Fig. 2

At the state just before calling the method `test`, the following conditions were obtained, i == \old(i) and j == 1. Since the method `test` has the first and the second arguments as x and y, the declaration of the method is `test`(i, j), and \old(x) and \old(y) correspond to i and j. Thus, the intermediate variables are $t_i and$t_j for i and j.

From the postcondition of `test` obtained by 3.3.1, the following final conditions are generated:

- $t_i == \old($i$)

- $t_j == 1
- ¬($t_i ≤ $t_j) ⟹ ret == $t_i − $t_j
- ($t_i ≤ $t_j) ⟹ ret == $t_j − $t_i

As the return value of the method `test` is assigned to ret, ret is substituted with \result.

3.3.6 Instances of Other Classes

This section describes the process for reference variables.

Assume that the reference variable b of class B is in the target method m of class A. Static analysis is performed on every field of class B. Further, assume that the field variables x, y, and z are in class B, while the target variables are $b.x$, $b.y$, and $b.z$.

Their values will be $b.x$ == \old($b.x$), $b.y$ == \old($b.y$), and $b.z$ == \old($b.z$) for the cases where variable b is an argument of method m or a field variable of class A. If the variable b is local, then apply the process mentioned in 3.3.5 to a constructor of b. The process will be recursive if some of x, y, and z are reference variables of other classes.

3.3.7 Arrays

This section describes the procedure for arrays

Assume that there is an array variable a in the target method. In general, it is not efficient for static analysis to be performed on each element of a.

However, static analysis is only performed when assignment to some element occurs. This approach is efficient when the number of assignment statements is small.

Case-splitting is performed based on the value of index. For example, a case of assignment "$a[0] = 10$," static analysis is performed first on "$a[0]$", while another case, "a[index] = 20", where the value of the variable "index" is greater than 0 is defined. Thus, both the case $index = 0$ and the other case must be considered. For the first case, analysis on $a[0]$ is performed, while for the second case, analysis on a[index] is performed.

The conditions "a[index] == 20" and "(index== 0) ⟹ $a[0] == 20$" are generated for a[index], while for $a[0]$ the conditions are "¬(index== 0) ⟹ $a[0] == 10$,".

In general, the following generation rule is used. For an array variable a, assume that the condition for its index index_{n-1} has already been obtained as follows:

$$\bigwedge_{i=0}^{m-1} (\text{cond}_i \Longrightarrow a[\text{index}_{n-1}] == \text{E}_i)$$

, where cond_i is a condition, and E_i is an expression.

When an assignment statement $a[\text{index}_n]$ =E is evaluated, the following condition for $a[\text{index}_{n-1}]$ is created:

$$\begin{cases} \bigwedge_{i=0}^{m-1}((\neg(\text{index}_n == \text{index}_{n-1}) \wedge \text{cond}_i) \\ \Longrightarrow a[\text{index}_{n-1}] == \text{E}_i) \\ (\text{index}_n == \text{index}_{n-1}) \Longrightarrow a[\text{index}_{n-1}] == \text{E} \end{cases}$$

3.4 Generation of Preconditions

There are several approaches to generate preconditions. One such approach is to enumerate as many of the states as is feasible where the target method is just called, and analyze the variables' values for each of the states.

Daikon uses many test cases and, thus, follows the above approach. However, in static analysis, it is difficult to follow such an approach, due to the complex calculations involved.

The proposed tool takes a simple approach by simply checking typical exceptions. In the symbol execution phase, if the tool finds a condition that might cause such exceptions, then it outputs the negation of such a condition as a precondition. The advantage of this method is that it is a lightweight analysis. The target exception is the null reference, and the array index is an out-of-bounds violation. These exceptions are the most common dangerous exceptions [10].

3.4.1 Null Reference

A null reference is a reference through a null-valued reference variable. JavaVM throws a NullPointerException for such a case.

For the method test(B, int) in Fig. 5, the null reference might occur when the value of b is null.

During the symbolic execution phase, the tool determines the possibility of a null reference. In this case, $i > 0 \Longrightarrow b$!= null is generated as a precondition.

```
class A {
  int num = -1;
  public int test(B b, int i) {
      int ret = 0;
      if(i > 0)
          ret = b.num;
      else
          ret = this.num;
      return ret;
  }
}

class B {
    static public int num = 10;
}
```

Figure 5: A Java method that may cause a null pointer exception

3.4.2 Array Index Out-of-bounds Violation

An array index out-of-bounds violation occurs when a reference occurs for an index outside the array region.

JavaVM throws an ArrayIndexOutOfBoundsException for such a violation.

If there is a reference to a[index] of array a, then 0 ≤ index < MAXLENGTH is generated as a precondition, where

MAXLENGTH is the size of the array usually given during its creation. If there is no creation statement, MAXLENGTH is $\backslash$old(a).length.

4 Experiment

This section provides evaluation experiments using the proposed tool and Daikon.

4.1 Outline

The proposed tool and Daikon were applied to the test(IntNum, IntNum) method of gnu.math.BitOps, which is included in the Kawa language framework[18]. Six different types of test cases were used.

This method calls the test(IntNum, int) method of the same class BitOps and the isNegative() method of gnu.math.IntNum. It uses operations and types that conform to the restrictions of the tool. It also uses arrays and other method calls, so that it can be used to test almost all of the functions in the proposed tool.

Table 2 shows the target method's metrics.

Table 2: Target Java Method's Metrics

Method	LOC	LOC (Jimple)
boolean test(IntNum, IntNum)	about 20	about 70
boolean test(IntNum, int)	about 10	about 50
boolean isNegative()	about 5	about 30

Different types of test cases were used to observe the reliability of the generated assertions. The coverage metrics are used to compare the test cases.

Table 3 shows the test cases and their coverage. jcoverage is used to measure the coverage.

Table 3: Coverage of Test Cases

test cases	coverage	test cases	coverage
TC1	55%	TC4	95%
TC2	70%	TC5	100%
TC3	90%	TC6	100%

TCi includes TCk $(k < i)$. TC6 adds more paths than TC5.

Detail of the test case are the following:

1. It executes the BitOps.test with 100 random values using every combination of int and long types.
2. It executes BitOps.test with random pairs of the same values and random pairs of negative values, as well as test case 1.
3. It executes BitOps.test with pairs consisting of $[0, 1]$, as well as test case 2.
4. It executes BitOps.test with random value setting of its array, as well as test case 3.
5. It executes BitOps.test with a random array size, as well as test case 4.

The reliability of generated assertions and execution times was evaluated using the proposed tool and Daikon.

4.2 Reliability of Assertions

The reliability of assertions was checked using the proposed tool and Daikon to evaluate the test cases described in 4.1. Table 4 shows the number of feasible (infeasible) assertions and the ratio between them and the number of all generated assertions.

For a given assertion, it was manually decided whether a given assertion is feasible. An assertion is said to be infeasible if the assertion is meaningless or wrong. It should be noted that not all assertions can be classified.

In [20], the ratio of the number of feasible assertions to the number of total assertions using only Daikon is stated to be more than 90% when using ESC/Java2 to evaluate the feasibility. However, ESC/Java2 can classify a meaningless assertion as feasible. For example, x[3]> x[1] for a sorting method is feasible with ESC/Java2 but meaningless; thus, based on the above criteria, the assertion is infeasible.

However, paper [19] shows that the recall and precision of the assertions generated by Daikon is less than 30% based on a manual calculation.

For every test case, about 80% of the generated assertions by Daikon are infeasible, while the generated tool generates far fewer infeasible assertions.

Filtering of assertions is a useful technique to reduce the number of infeasible assertions. However, it also reduces the number of final generated assertions.

The result of $\langle 1 \rangle$ suggests the proposed tool generates more feasible assertions than Daikon, which is an encouraging result.

The number of assertions produced by Daikon varies with test case type, while the proposed tool produces similar results irrespective of the test case type. This helps the user to provide suitable test cases, since test cases that apply only to loop sections can be selected. Such test cases will improve the reliability of generated assertions. However, this approach imposes more workload on users. Thus, there are trade-offs between the reliability of assertions and the manual work required.

In this experiment, the same test cases were used, and the results show that this is sufficient.

4.3 Execution Times

The execution times of the proposed tool and Daikon were measured using the test cases described in 4.1.

The execution time of Daikon is the "execution time of Daikon applied to the target method," while that of the proposed tool is the sum of the "execution time for the Prepara-

Table 4: Comparison of Daikon and the Proposed Tool

Test Cases	Daikon					Our tool				
	⟨1⟩		⟨2⟩		Sum.	⟨1⟩		⟨2⟩		Sum.
1	0	0%	19	95%	20	18	72%	1	4%	25
2	9	10%	73	85%	86	18	72%	1	4%	25
3	10	11%	76	87%	87	27	71%	0	0%	38
4	18	7%	201	82%	245	27	71%	0	0%	38
5	20	15%	110	82%	134	27	71%	0	0%	38
6	21	22%	74	78%	95	27	71%	0	0%	38

⟨1⟩: The number of feasible assertions (ratio between feasible and total assertions)

⟨2⟩: The number of infeasible assertions (ratio between infeasible and total assertions)

tion stage, the execution time for Daikon applied to the loop sections, and the execution time for symbolic analysis."

In general, the execution time of Daikon is linear to the product of the number of program points and the number of test cases (the number of executions).

Table 5 shows the results.

Table 5: Execution Times

	Daikon (sec)	Our tool (sec)
TC1	17.7	39.3
TC2	32.2	39.7
TC3	88.3	40.8
TC4	100.4	41.1
TC5	115.4	41.1
TC6	149.0	41.2

For TC3, 4, 5, and 6, the proposed tool runs faster than Daikon. This can be explained as follows. Daikon calls BitOps.test(IntNum, int) and IntNum.isNegative() twice and once, respectively in the target method BitOps.test(IntNum, IntNum). Thus, the number of program points becomes 12 $(2+(2+2)\times 2+2)$. On the other hand, in the proposed tool, because the number of loop sections is one, the number of program points is two. Thus, Daikon's single path approach leads to a longer execution time.

For TC1 and 2, Daikon runs faster than the proposed tool. However, table 4 shows that more infeasible assertions are generated with Daikon than with the proposed tool. This negates the gain in time that Daikon has.

For TC5 and 6 with 100% coverage, the proposed tool is more useful. Other experiments show that the execution time of Daikon does not increase when the program size increase. However, the proposed tool's execution time increases with program size.

Therefore, scalability of the proposed tool is a drawback. However, since, in practice, the proposed tool can be applied to each method independently, this is not a major drawback.

5 Conclusions

In order to increase the reliability of assertions, this paper proposed a hybrid method for generation of pre- and postconditions for Java programs, where Daikon is only used for the loop sections, and static analysis is used for the other sections.

The experimental results show that the number of false assertions generated by the proposed method is less than that using Daikon alone.

Future work will include the generation of class invariants. As well, a new hybrid approach to generating assertions that combines symbolic execution and model checking techniques will be developed.

REFERENCES

[1] B. Meyer, "Applying Design by Contract," in Computer, vol.25, no.10, pp.40–51, 1992.

[2] B. Meyer, "Eiffel: The Language," Prentice Hall, 1991.

[3] G.T. Leavens, A.L. Baker, and C. Ruby, "JML: A notation for detailed design," In Behavioral Specifications of Businesses and Systems, pp.175–188, Kluwer Academic Publishers, 1999.

[4] G.T. Leavens, A.L. Baker, and C. Ruby, "Preliminary design of JML: A behavioral interface specification language for Java," Technical Report 98–06u, Iowa State University, Dept. of Computer Science, 2003.

[5] Y. Cheon and G.T. Leavens, "A Runtime Assertion Checker for the Java Modeling Language(JML)," Int. Conf. on Software Engineering Research and Practice, pp.322–328, 2002.

[6] C. Flanagan, K.R.M. Leino, M. Lillibridge, G. Nelson, J.B. Saxe, and R. Stata, "Extended static checking for Java," Proc. of the ACM SIGPLAN 2002 Conf. on Programming language design and implementation, pp.234–245, 2002.

[7] J.van den Berg and B. Jacobs, "The LOOP compiler for Java and JML," Proc. of the 7th Int. Conf. on Tools and Algorithms for the Construction and Analysis of Systems, pp.299–312, 2001.

[8] R.E. Rodriguez, M.B. Dwyer, and J. Hatcliff, "Checking Strong Specifications Using An Extensible Software Model Checking Framework," Proc. of the 10th Int.

Conf. on Tools and Algorithms for the Construction and Analysis of Systems, pp.404–420, 2004.

[9] L. Burdy, Y. Cheon, D. Cok, M.D. Ernst, J. Kiniry, G.T. Leavens, K.R.M. Leino, and E. Poll, "An overview of JML tools and applications," Int. Journal on Software Tools for Technology Transfer, vol.7, no.3, pp.212–232, 2005.

[10] C. Flanagan and K.R.M. Leino, "Houdini, an Annotation Assistant for ESC/java," Proc. of the Int. Symp. of Formal Methods Europe on Formal Methods for Increasing Software Productivity, vol.2021, pp.500–517, 2001.

[11] G. Ammons, R. Bodk, and J.R. Larus, "Mining specifications," Proc. of the 29th Annual ACM SIGPLAN-SIGACT Symp. on Principles of Programming Languages, pp.4–16, 2002.

[12] P. H. Schmitt and B. Weiß, "Inferring Invariants by Symbolic Execution," Proc. of the 4th Int. Verification Workshop (VERIFY 2007), pp.195–210, 2007.

[13] M.D. Ernst, "Dynamically Discovering Likely Program Invariants," Ph.D thesis, Univ. of Washington Dep. of Computer Science and Engineering, 2000.

[14] M.D. Ernst, J. Cockrell, W.G. Griswold, and D. Notkin, "Dynamically discovering likely program invariants to support program evolution." IEEE Trans. Software Engineering, vol.27, no.2, pp.1–25, 2001.

[15] S.J. Garland, N.A. Lynch, and M. Vaziri, "IOA: A language for specifying, programming, and validating distributed systems," Technical Report, MIT Laboratory for Computer Science, 1997.

[16] N. Dodoo, L. Lin, and M.D. Ernst, "Selecting, refining, and evaluating predicates for program analysis," Technical Report, MIT Laboratory for Computer Science, MIT-LCS-TR-914, 2003.

[17] R. Vallee-Rai, L. Hendren, V. Sundaresan, P. Lam, E. Gagnon, and P. Co, "Soot - a java optimization framework," Proc. of the 1999 Conf. of the Centre for Advanced Studies on Collaborative Research, pp.125–135, 1999.

[18] Kawa language framework, http://www.gnu.org/software/kawa/index.html

[19] K. Miyamoto, K. Okano, and S. Kusumoto, "Adequacy evaluation of assertions for Java including loop invariants generated by Daikon," The IEICE Trans. on Information and System, Vol.J91-D, No.11, to appear, 2008 (in Japanses).

[20] J. W. Nimmer and M. D. Ernst, "Automatic genera- tion of program specifications," in Proc. of the 2002 Int. Symp. on Software Testing and analysis (ISSTA), pp. 232–242, 2002.

Modeling Language for LDAP and Automatic Production System

Kenji Saotome[*], Seiichi Kondo[*], Akihisa Oonuma[*],
Takashi Komiya[*], Masahiko Ishino[*], Tsukasa Kudo[*],
Sanshiro Sakai[*] and Tadanori Mizuno[*]

[*]Hosei Business School of Innovation Management, Japan
[*] Mitsubishi Electric Corporation, Information Technology R&D Center
[*] Mitsubishi Electric Information Technology Corporation Japan
[*] Mitsubishi Electric Information Systems Corporation, Japan
[*] Faculty of Informatics, Shizuoka University, Japan
[*] Graduate School of Science and Technology, Shizuoka University, Japan

Abstract - LDAP Directory Service begins to be used as a tool for the development of Enterprise Information System. Nevertheless, there are not the standards of the model for the design of the Directory and generally incomplete diagrams of the Directory have been illustrated. The methods to design the Directory are expected. Then, we proposed the modeling language extending UML for the design of Directory Information Tree (DIT). We developed for trial the system that automatically generates the programs that manage the Directory, and evaluated it. We found that this system is enough applicable and efficient.
Keywords: Directory, LDAP, UML, MDA, Modeling.

1 INTRODUCTION

In late years the standardization of the directory service advances, and the directory service products which manage the data of the information system of the company come to be released, and it begins to be used. However, when the needs to build the original data structure by the system matter occurred, it becomes custom to show the incomplete figure of the degree to exemplify hierarchical structure and do it with a design document because there is not a standard design model peculiar to the directory. Therefore by the conventional construction technique of the directory, it becomes difficult to get the desired directory for the program developer and the user because we are not able to realize the smooth mutual understanding between the directory designer and the program developer and the user using it. There is the danger that a problem occurs just before the operation after the development.

We already proposed *Directory Modeling Language* specialized in the directory service based on the UML model, and produced the system(*Automatic Production System*) which generated *Directory Management Program* from this modeling language automatically and demonstrated the applicability of this system [1].

This research domain is the directory, programs automatic production and UML expansion, but the other researches corresponding to these all domains are not found. Therefore the theme of this research is an advanced one [2][3] [4].

In this time, we added the notation of the schema definition of the object class and the attribute type to the modeling language and realized the automatic generation function of the schema, and the system implementing the specifications of the practical use level was completed. By this report, we report about Directory Modeling Language and the specifications of Directory Management Program that Automatic Production System generates and the evaluation of this system.

The directory of this report has the structure of ITU-T (International Telecommunication Union-Telecommunication Standardization Sector) Recommendation X.500 series and the interface of LDAP(Lightweight Directory Access Protocol) defined by IETF(Internet Engineering Task Force)[5][6][7]. The directory modeling language to propose is based on UML(Unified Modeling Language) standardized in OMG(Object Management Group) [8][9].

2 MODELING LANGUAGE SPECS

This section describes the syntax specifications and the semantic specifications of Directory Modeling Language.

2.1 Syntax Specification

The model of Directory Modeling Language is expressed by the *class diagram* of UML. In the class, the *stereotype* and the *tagged value* is specified to show the role of the class. We show below the syntax specifications of this modeling language..

2.1.1 <<LDAP>> Class

(1)For a *model*, there must be only one *class* specifying an ≪LDAP≫stereotype.

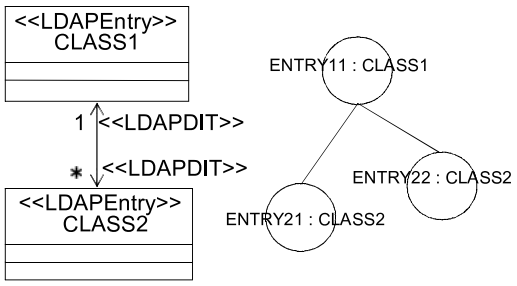
(2)In this class, a {LDAPRoot}tagged value must be specified. In {LDAPRoot} tagged value, a *character string* is specified.
(3)In this class, the *association* must not be specified.

2.1.2 <<LDAPEntry>> Class

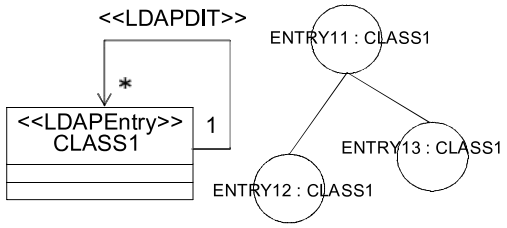
(1)In this model the class name must be unique among the classes specified <LDAPEntry> stereotype.
(2)In the class, one or more {LDAPObjectClass}tagged values must be specified. In {LDAPObjectClass} tagged value, a character string is specified.
(3)The class has one or more *attributes*.
(4)The attribute consists of the attribute name and the *type*. The attribute name must be unique in this class.
(5)The type must be **String**, **byte[]**, **Collection**, **Collection<String>** or **Collection<byte[]>**.
(6)<LDAPRDN>stereotype must be specified for one or more attributes in a class.
(7) The association between classes is able to be specified, and the *multiplicity*, the *navigability* and the *role* name at the association end are able to be specified. When there are one or more associations in a class, the role name at each association end must be unique each other.
(8)The multiplicity of the association is specified the following one. The notation (*) means more than 0.
1
*
(9) The arrow of both directions or the arrow of single direction is specified for the navigability of the association.
(10) At the association end, one stereotype of the following kinds must be specified.
<LDAPDIT>
<LDAPDN>
<LDAPAttr>
(11)When <LDAPDIT>stereotype at the association end is specified, <LDAPDIT>stereotype must be specified at another association end if nessesary. The multiplicity at this association must be *vs. 1.
(12) At the association end of <<LDAPAttr> stereotype, {LDAPKey} tagged value must be specified.

2.1.3 <<LDAPDefAttributeTypeS>> Class

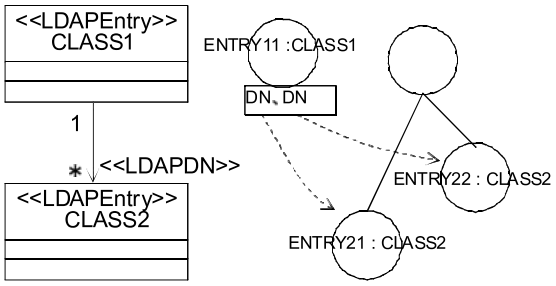
(1)In this class, a {LDAPSyntax}tagged value must be specified. A character string is specified in {LDAPSyntax}tagged value.
(2)The class has one or more attributes.
(3)The attribute consists of the name and the type.
(4)The attribute name must be unique among the attribute names described in the classes of <LDAPDefAttributeTypeS> stereotype.
(5)In the type, **void** or **Collection** is specified .



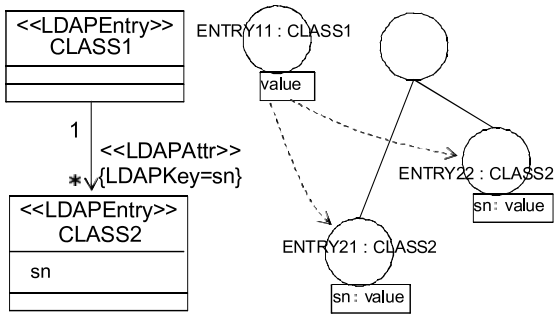
(a) <LDAPDIT>and DIT.



(b) <<LDAPDIT>and DIT(Self Assoc).



(c) <LDAPDN>and DIT.



(d) <LDAPAttr>and DIT.

Figure 1 : Modeling Language and DIT.

(6)In this attribute, if necessary, {LDAPEquality}, {LDAPOrdering} or {LDAPSubstr} tagged value is specified.
(7)In this class, the association must not be specified.

2.1.4 <<LDAPDefObjectClass>> Class

(1)In this class, a {LDAPSuperior}tagged value must be specified. In {LDAPSuperior} tagged value, a character string is specified.
(2)A class has one or more attributes.
(3)The attribute consists of the attribute name and the type.
(4)The attribute name must be unique in a class.
(5)**void** must be specified if specified.
(6)For each attribute, «LDAPMust>stereotype is able to be specified.
(7)In this class, the association must not be specified.

2.2 Semantic Specifications

2.2.1 Class

2.2.1.1 <<LDAP>> Class

In the class specified «LDAP>stereotype, the information of the whole model is specified.

In {LDAPRoot}tagged value specified in this class, *DN(Distingushed Name)* which is top entry of all directory entries belonging to the classes of this model is specified.

2.2.1.2 <<LDAPEntry>> Class

The class specified «LDAPEntry> stereotype defines the directory entry. In the class, an *object class* of the directory constituting an entry belonging to the class is specified in the {LDAPObjectClass} tagged value. One or more {LDAPObjectClass}tagged values are able to be specified, but **top** object class is omitted.

The each attribute name defined in the class must be the *attribute type* of the directory to belong to one of the object classes specified in {LDAPObjectClass}tagged value.

The type of the attribute defined in the class is the type of JAVA used in APIs which Automatic Generation System generates.

When the *attribute syntax* shows binary data such as **Binary**, **Octet String** or **Certificate**, **byte[]** is specified when a single value, and **Collection<byte[]>** is specified when multi values. When the attribute syntax shows string data such as **Directory String**, **Boolean** or **Integer**, string is specified when a single value, and **Collection** or **Collection<String>** is specified when multi values.

«LDAPRDN> stereotype shows that it is an attribute type to become *RDN (Relative Distinguished Name)* of the entry.

2.2.1.3 <<LDAPDefAttributeTypeS>> Class

The class specified «LDAPDefAttributeTypeS> stereotype defines the user-defined attribute type with the *attribute syntax*. The attribute syntax is specified in {LDAPSyntax}tagged value of the class. The attribute syntax (the **SYNTAX** keyword in the attribute type definition of the directory) is specified in {LDAPSyntax}tagged value.

The attribute of the class shows the attribute type of the directory, and the attribute name shows the name of attribute type. In the type, **void** must be specified when attribute type of single value, and **Collection** must be specified when attribute type of multi values.

If necessary, {LDAPEquality}, {LDAPOrdering}and {LDAPESubstr}tagged value are able to be specified in each attribute of the class. {LDAPEquality}, {LDAPOrdering} and {LDAPESubstr} tagged value show the *matching rule* of **EQUALITY** , **ORDERRING** and **SUBSTR** each.

For the character string to specify in these tagged values, the keyword of the matching rule of the attribute type definition of the directory must be specified.

2.2.1.4 << LDAPDefObjectClass >> Class

The class specified «LDAPDefObjectClass> stereotype defines user-defined object class. In {LDAPSuperior}tagged value specified in the class, the superior object class (the keyword **SUP** of the object class definition) is specified.

The attribute of the class shows the attribute type to belong to this object class. The attribute name shows the attribute type name. «LDAPMust>stereotype in each attribute shows that this attribute type is the required attribute type of this object class.

2.2.2 Association

The association shows the association between the entries. Either of «LDAPDIT>, «LDAPDN>or «LDAPAttr> stereotype is specified at the association end to show the implementation of the association of the directory.

2.2.2.1 Association of <<LDAPDIT>>

«LDAPDIT>stereotype shows that the association is implemented by *DIT (Directory Information Tree)*. The multiplicity of this association must be *vs. 1. This association is implemented by the method that an entry belonging to the class of the association end specified "1" is posted as the direct upper entry of the entry belonging to the class at the other association end.

When there are more than 2 associations by «LDAPDIT>stereotype in a class, this class cannot have more than 2 association that have the "*" multiplicity among these association ends at this class side. Figure 1(a) shows an example of the class to use «LDAPDIT>stereotype and DIT implemented. The implementation of the association by DIT is the most natural structure for the directory. As to express the organization of the company, the self association to

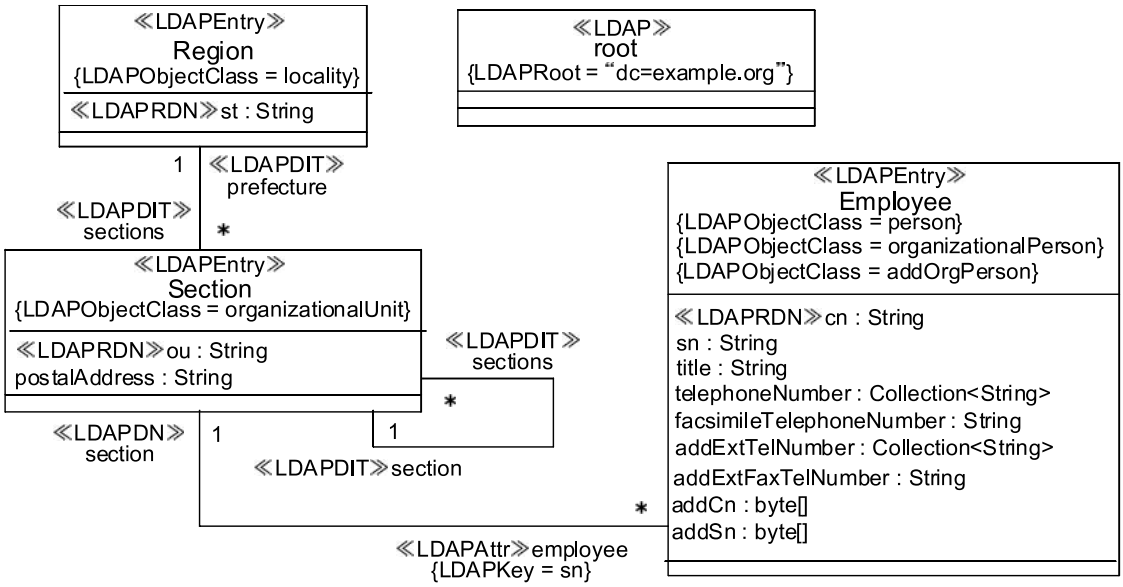


Figure 2 : DIT Model by Directory Modeling Language.

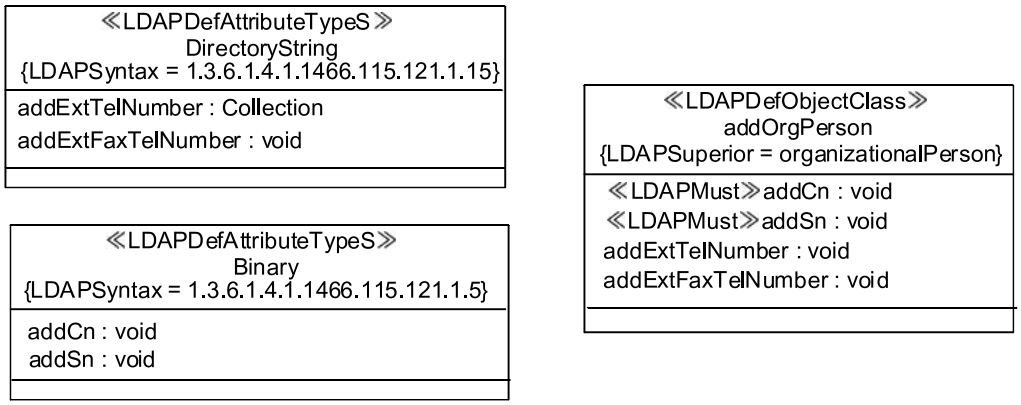


Figure 3 : Scheme Definition Model by Directory Modeling Language

associate with the own class is the typical example associated by DIT of the directory. Figure 1(b) shows the class diagram expressing the self association and DIT implemented.

2.2.2.2 Association of <<LDAPDN>>

The ≪LDAPDN≫ stereotype shows that the association is implemented by DN. An entry belonging to a class has DN of an entry belonging to the class at the other association end. This implementation resembles the method that is used for the relations of the group and the member to be called the static group in the directory. Figure 1(c) shows the class diagram to use the ≪LDAPDN≫ stereotype and DIT implemented. The dotted lines in this figure are the notation to show the associated entries, and are not to constitute the real DIT.

2.2.2.3 Association of <<LDAPAttr>>

The ≪LDAPAttr≫ stereotype shows that the association is implemented using an attribute value of the attribute type. {LDAPKey} tagged value is specified on the association end. In {LDAPKey} tagged value, the attribute name to use for the association among the attributes of the class at the other association end is specified. When the value for this association in the entry is equal to the attribute value in the attribute type specified by the {LDAPKey} tagged value belonging to the class of the other association end, it is considered that these entries are associated. This implementation resembles the relations of the primary key and the foreign key of the relational data base. Figure 1(d) is the figure of the class diagram specified ≪LDAPAttr≫ stereotype and DIT implemented.

```
Region {
   #entry (st="Tokyo") {
      Section {
         #entry (ou="Sales", postalAddress="Shinjuku-ku, Tokyo", employees=<sn="A"> <sn="B">) ;
         #entry (ou="development", postalAddress="1, Chiyoda-ku, Tokyo", employees=<sn="C">) {
            Section {
               #entry (ou="development-1", postalAddress="2, Chiyoda-ku, Tokyo", employees=<sn="D"> <sn="E">) ;
               #entry (ou="development-2", postalAddress="3, Chiyoda-ku, Tokyo", employees=<sn="F"> <sn="G">) ;
 ……..……
}
Employee {
   #entry (cn="A", sn="A", title="chief", telephoneNumber="03-1111-0001", facsimileTelephoneNumber="03-1111-1001",
      addCn="gmA=", addSn="gmA=", addExtTelNumber="11-0001", addExtFaxTelNumber="11-1001", section=<ou="Sales">) ;
   #entry (cn="B", sn="B", title="staff", telephoneNumber="03-1111-0002",  facsimileTelephoneNumber="03-1111-1002",
      addCn="gmE=", addSn="gmE=", addExtTelNumber="11-0002", addExtFaxTelNumber="11-1002", section=<ou="Sales"> ) ;
 ……..……..
   #entry (cn="I", sn="I", title="staff", telephoneNumber="06-2222-0002",  facsimileTelephoneNumber="03-2222-1002",
      addCn="gmg=", addSn="gmg=", addExtTelNumber="22-0002", addExtFaxTelNumber="22-1002", section=<ou="Personnel"> ) ;
}
```

Figure 4 : Input of Directory Loading Compiler.

3 DIRECTORY MANAGEMENT PROG

3.1 Outline

We developed for trial the system(Automatic Production System) to generate automatically the program(Directory Management Program) to manage the directory from the model of Directory Modeling Language.

Directory Management Program has the following specifications so that assuming trial.

- On ly import by batch of all data
- Only access by the application program

Directory Management Program consists of *Directory Loading Compiler*, *Directory Access API* and *User-defined Schema Information*.

The example of figure 2 assumes the directory which manages the information of the organization of the company, the region where it is located at and the employee who belonged to it. We select the standard attribute type as much as possible, but define the user-defined attribute type when we cannot use it. In figure 2, **addExtTelNumber**, **addExtFaxTelNumber**, **addCn** and **addSn** of **Employee** are the user-defined attribute type. **addExtTelNumber** and **addExtFaxTelNumber** show the extension telephone or FAX number. **addCn** and **addSn** are additional **Cn** and **Sn**. For example these are defined for the purpose of keeping the full name that includes GAIJI with the character code set (Shift-JIS etc) except utf-8.

Figure 3 describes the definition information of these user-defined attribute types. **addExtTelNumber** and **addExtFaxTelNumber** are defined as the attribute type of the ⊱DAPDefAttribu teTypeS> class which name is **DirectoryString**. **addCn** and **addSn** are defined as **Binary** because they are expressed with the character code set except utf-8. The **addOrgPerson** object class that these user-defined attribute types belongs to is defined in the ⊱DAPDefObjectClass≻c lass.

SectionHome
SectionHome (ctx: DirContext) : SectionHome findByPrimaryKey (ou: String) : Section findAll () : Collection

Section
Section (ctx: DirContext, dn: String, ou: String, postalAddress: String, employees: Collection) : Section getDn () : String getOu () : String getPostalAddress () : String getRegion () : Region getSection () : Section getSections () : Collection getEmployee () : Collection

SectionDTO
SectionDTO (ou: String, postalAddress: String) : SectionDTO getOu () : String setOu (ou: String) : void getPostalAddress () : String setPostalAddress (postalAddress: String) : void

Figure 5 : Methods of Class "Section".

3.2 Directory Loading Compiler

3.2.1 Characteristics

Directory Loading Compiler has the following characteristics.

(1)Incomplete information is not left in the directory because the loading script can detect the structural error of DIT at the compilation time.

(2)The loading script can check the relation of definition and reference about the association.

(3)To use the LDIF form it is necessary to be specified the DN of the each entry itself and the related entry, but to use the loading script the loading data can be specified without being conscious of DN

3.2.2 Specifications

Directory Loading Compiler interprets the loading script for the loading directory data by batch, and generates the loading data of the LDIF form. Figure 4 is the script described in the specifications of Directory Loading Compiler generated by the model of Figure 2. **Region**, **Section** and **Employee** of Figure 4 is the class name, and the entry corresponding to this each class is specified in **#entry()**. In **#entry()**, the attribute type and the attribute value is specified in the form of **parameter=value**. For example, **st="Tokyo"** of the entry of **Region** shows that **st** is the attribute type and **Tokyo** is the attribute value.

At the following {} of **#entry()**, the entry being the lower entry in DIT is specified. The entry of **ou="sales department"** and **ou="Development Division"** of **Section** class becomes a direct lower entry which RDN is **st="Tokyo"**. In this way, the DN of each entry is decided by expressing DIT structure with the nest and RDN of each entry.

The role name in the class diagram is used to express the association except DIT. The **section** parameter specified in the entry of **Employee** class is the role name specified at the ŁDA PDN> stereotype of the class diagram, and the associated entry at this parameter is specified.

3.3 Directory Access API

3.3.1 Characteristics

Directory Access API has the following the characteristics.

(1)The API uses JNDI (JAVA Naming and Directory Interface) which is JAVA standard interface of LDAP directory access and is implemented as the API which does not depend on the directory server.

(2)Directory Access API can be accessed with the interface of the design pattern adopted in EJB (Enterprise JavaBeans).

```
dn: cn=schema
changetype: modify
add: attributetypes
attributeTypes: ( addExtTelNumber-oid
 NAME 'addExtTelNumber'
 DESC 'User Defined Attribute'
 SYNTAX 1.3.6.1.4.1.1466.115.121.1.15
 X-ORIGIN 'user defined' )
attributeTypes: ( addExtFaxTelNumber-oid
 NAME 'addExtFaxTelNumber'
 DESC 'User Defined Attribute'
 SYNTAX 1.3.6.1.4.1.1466.115.121.1.15
 SINGLE-VALUE
 X-ORIGIN 'user defined' )
attributeTypes: ( addCn-oid
 NAME 'addCn'
 DESC 'User Defined Attribute'
 SYNTAX 1.3.6.1.4.1.1466.115.121.1.5
 SINGLE-VALUE
 X-ORIGIN 'user defined' )
attributeTypes: ( addSn-oid
 NAME 'addSn'
 DESC 'User Defined Attribute'
 SYNTAX 1.3.6.1.4.1.1466.115.121.1.5
 SINGLE-VALUE
 X-ORIGIN 'user defined' )
```

Figure 6 : LDIF for Definition of Attribute Types.

```
dn: cn=schema
changetype: modify
add: objectclasses
objectClasses: ( addOrgPerson-oid
 NAME 'addOrgPerson'
 SUP organizationalPerson STRUCTURAL
 MUST ( addCn $ addSn )
 MAY ( addExtTelNumber
 $ addExtFaxTelNumber )
 X-ORIGIN 'user defined' )
```

Figure 7 : LDIF for Definition of Object classes.

3.3.2 Specifications

Directory Access API has the functions to get the entry, the attribute value of the entry and the associated entry to the entry and so on.

Figure 5 shows the function of the API to access the entry belonging to **Section** class of Figure 2. **findByPrimaryKey()** method of **SectionHome** class gets the instance of **Section** class by the attribute type of RDN. **findAll()** method get all the entries belonging to **Section** class. **getDn()**, **getOu()** and **getPostalAddress()** of **Section** class get the attribute value of the entry. **getRegion()**, **getSection()**, **getSections()** and **getEmployee()** get the associated entry. **SectionDTO** class is the JavaBeans implemented **Java.io.Serializable** and has all of the attribute values.

3.4 User-defined Schema Information

3.4.1 Characteristics

User-defined Schema Information has the following characteristics.

(1)The User-definition Schema Information does not depend on the directory server because it is generated by the standard LDIF form.

3.4.2 Specifications

The User-defined Schema Information consists of the data of the LDIF form to register the user-defined attribute type generated from the information of the class of «DAPDefAttributeTypeS» stereotype and the user-defined object class generated from the information of the class of <<LDAPDefObjectClass>> stereotype. Figure 6 is the LDIF formed data of the attribute type definition generated from the «DAPDefAttributeTypeS» class of Figure 3. Figure 7 is the LDIF formed data of the object class definition generated from the «DAPDefObjectClass» class of Figure 3.

4 EVALUATION

In comparison with the existing application program using the directory, we constructed the directory, tested the function and measured the performance.

The object of the evaluation is Information Leakage Prevention Solution of Mitsubishi Electric Corporation (Mitsubishi Solution System as follows)[10]. We describe the functionally equal directory with Directory Modeling Language for the organization and the employee information of the directory of Mitsubishi Solution System, compare the performance of the each program that we make with Directory Access API and the API of Mitsubishi Solution System.

4.1 User-defined Schema definition

We defined 35 user-defined object classes and 136 attribute types of Mitsubishi Solution System and were able to confirm the same schema definition.

By this definition, the classes in the model that we described by Directory Modeling Language were 7 «DAPDefAttributeTypeS» classes and 35 «DAPDefObjectClass» classes.

What 136 attribute types are able to be expressed by 7 «DAPDefAttributeTypeS» classes shows that the attribute syntaxes of these attribute types are 7 kinds.

4.2 Building DIT

We were able to confirm that the DIT approximately same as Mitsubishi Solution System was build.

Table 1 : Data Structure of Measurement

organization	org units	members	Title
	1	1	president
head office	5	5	vice president
division	25	25	general manager
deaprtment	125	125	director
section	625	5,000	manager, staff
TOTAL	781	5,156	

Table 2 : Result of Measurement

Program	system	time (minutes)	num of entries sent to PC
Program1	MMS	20	19,842
	DMP	24	19,061
Program2	MMS	54	40,081
	DMP	50	40,000

There were 9 «DAPEntry» classes, 30 attributes and 13 associations in the model that we described.

4.3 Performance Measurement

4.3.1 Measurement Method

4.3.1.1 Data Structure

We assume the company of 5,000 employees and compare the performance using data for the measurement such as Table.1.

4.3.1.2 Evaluation Programs

We assume the implementation of the address book by the directory and use the following two programs for the performance comparison.

(1) Program1

This is the program that is assumed to find the employees from the organization hierarchy. While referring in sequence from the main office to the lower organization, this program displays the data of the organizational unit and all the employees belonging to there. The entries to access are 781 organizational units and 5156 employees.

(2) Program2

This is the program that is assumed to directly find an employee by the attribute. This program generates an employee name at random and acquires the data of the employee with the employee name as a key. The entries to access are 5000 employees, 5000 organizational units and 4,999 upper heads except the president.

4.3.1.3 Measurement environment

We connect two following computers in 100BaseT and access from PC that the evaluation programs execute to the directory server in LDAP interface.

(1)The PC that the evaluation programs execute
H/W CPU: Intel Pentium4 2.8GHz
Memory: 760MB, HDD: 35GB
S/W Java 1.4.2 06
Windows XP Professional
(2)The directory server
H/W CPU: Intel Xeon 3.2GHz
Memory: 2GB, HDD: 292GB
S/W SunONE Directory Server 5.2
Windows Server 2003

4.3.2 Result

Table 2 shows the measurement result. In Table 2, MMS is Mitsubishi Solution System, and DMP is Directory Management Program.

4.3.3 Discussion

We found that Directory Modeling Language has enough description ability of the application system, because with Directory Modeling Language, we was able to build DIT approximately same as Mitsubishi Solution System.

We found that Directory Access API is implemented by approximately same method as the API of Mitsubishi Solution System in the access of the directory, because about the number of the entries sent to PC, the difference for program1 is 781 and the difference for program2 is only 81.

We found that Directory Access API can work with practical performance, because the total of the execution times of Program1 and Program2 is same as Mitsubishi Solution System.

5 CONCLUSION

We found that this system is enough applicable and efficient. By using the result of this work, the improvement of the quality, the productivity and the maintenancibility can be expected.

We will implement OCL(Object Constraint Language) for the expansion of Directory Modeling Language in the future[11]. OCL is the function taken for the standard in UML2.0 and the language to describe the limitation and the query of the UML model.

Now we have the most basic function of the acquisition of the entry such as the follows.

- **findByPrimaryKey()**
  find the entry by the primary key(RDN) in the class.
- **findAll()**
  find the all entry in the class.

When the search by the free search condition is necessary, it assumes that user oneself learn the implementation of the Directory Access API and realize it. In the application system, it is necessary to acquire the entry by the complex search condition of the attribute value etc. We will implement the search with the free search condition by OCL.

REFERENCES

[1] K.Saotome, S.Kondo, A.Oonuma, T.Komiya, S.Sakai, T.Mizuno : ”Modeling Language for LDAP and Automatic Production System of Directory Management Program”, Information Processing Society of Japan, Database, Vol47, No.SIG13(TOD31), pp.28-39 (2006).

[2] Dragan Milicev : ”Automatic Model Transformations Using Extended UML Object Diagrams in Modeling Environments”, IEEE TRANSACTIONS ON SOFTWEAR ENGINEERING VOL.28 NO4, pp.413-431, 2002.

[3] Ludovic Apvrille, Jean-Pierre Court : ”TURTLE A Real-Time UML Profile Supported by a Formal Validation Toolkit”, IEEE TRANSACTIONS ON SOFTWEAR ENGINEERING VOL.30 NO7, pp.473-487, 2004.

[4] MDA(Model Driven Architecture),
UML Profile for EDOC, UML Profile for EAI.
http://www.omg.org/mda/specs.htm#Profiles

[5] ITU-T(International Telecommunication Union-Telecommunication Standardization Sector) .
http://www.itu.int/ITU-T/

[6] IETF(Internet Engineering Task Force) .
http://www.ietf.org/

[7] Tim Howes : ”Understanding and Deploying Ldap Directory Services”, Macmillan Technical Pub, 2003.

[8] OMG(Object Management Group).
http://www.omg.org/

[9] UML(Unified Modeling Language).
http://www.uml.org/

[10] Information Leakage Prevention Solution of Mitsubishi Electric Corporation.
http://global.mitsubishielectric.com/bu/security/rd/rd0702.html

[11] OCL(Object Constraint Language).
http://www.omg.org/technology/documents/formal/ocl.htm

Stepwise Approach to Design of Real-Time Systems based UML/OCL with Formal Verification

Takeshi NAGAOKA†, Eigo NAGAI†, Kozo OKANO†, and Shinji KUSUMOTO†

†Graduate School of Information Science and Technology, Osaka University
Machikane-yama 1–3, Toyonaka City, Osaka, 560–8531 Japan

Abstract - This paper provides a stepwise method for the design of real-time systems with timeliness QoS guarantees. In the proposed method, the system components are designed using UML diagrams and are provided with the timeliness QoS annotated with OCL. The basis of this technique is to formally ensure that the required timeliness QoS is satisfied under the provided timeliness QoS, given the network property and the UML diagrams. In order to avoid the state-explosion problem during performance model checking, which can logically check the satisfiability, the problem is separated into two steps. The first step checks the satisfiability using an abstract model of each of the components derived automatically from the provided QoS. The second step independently performs model checking for each of the components using a more detailed version of the behavioral model of a given component. Such an approach reduces the number of total states to check. Furthermore, the approach can be extended into hierarchical design, which leads to good scalability. Experimental results are also included in this paper.

Keywords: timeliness QoS, UML/OCL, model checking, component based systems, hierarchical design

1 Introduction

This paper presents a new method to verify consistency of timeliness QoS of component-based designed real-time systems. We assume that timeliness QoS is not only given to a whole system (Required QoS) but also associated with each component of a given system (Provided QoS).

Timeliness QoS is a time aspect of QoS (Quality of Service) features[1]. In the paper, we treat jitter, latency and throughput as timeliness QoS.

Nowadays, most real-time systems are designed with help of UML diagrams[7]. Especially components and their relation through signal communication can be represented in a class diagram of UML. In UML based design, such timeliness QoS can be annotated in OCL[8]. The annotation is associated to each of components as a provided QoS and also to a network link as a network property. Recently, SysML (System Model Language)[13] also attracts interest. SysML extends from UML and presents mixed systems consisting of physical devices and software and network systems. Therefore, SysML supports diagrams for signal flows and physical flows. For behavioral diagrams, SysML supports four diagrams as same as UML. Among them, state diagram has powerful representation including parallel and hierarchical states.

The proposed method is revised version of paper [15]. The method in [15] uses Linear Programming (LP) for some of verification. The approach has a disadvantage that connection among components has to be acyclic, and it cannot be applied to hierarchical design. The method of this paper uses abstract QoS automata instead of using LP; thus it improves the former disadvantage.

The heart of the technique is formally to ensure that the required timeliness QoS is satisfied under the provided timeliness QoS, given network property and the class diagram.

In order to check the satisfiability, there are several approaches. Formal approaches are very useful. Model checking is one of such an approaches. Notion of Test Automata[3], [5] and its application is also useful. However, one of disadvantage of the method is the state-explosion problem.

In order to avoid state-explosion problem while performing model checking, we separate the problem into two steps. The first step checks the satisfiability using abstract model of each of components derived automatically from the provided QoS. The second step performs model checking each of components independently using more detailed version of behavioral model of a component. Such an approach efficiently reduces the number of total states to check. Moreover the approach can be extended into hierarchical design; therefore it has good scalability.

SaveCCM[14] is a technique for Component based Development (CBD). In a description of a component, it allows user to define ports where signals input or output and to represent behavior in a timed automaton[2]. An IDE over Eclipse is available. Therefore, our proposed method has an affinity for SaveCCM.

The paper organized as follow. Section 2 gives timeliness QoS. Section 3 shows how to design component based real-timed systems in UML/OCL. Section 4 demonstrates the proposed method which consists of two steps. Section 5 provides an experimental example. We conclude the paper in Section 6.

2 Timeliness QoS

Main building blocks of our model are components. Each component has one or more *interfaces* to the environment, where all interactions between components is conducted via the interfaces. Since, we are mostly dealing with real-time systems and timeliness QoS, we shall assume that the interaction of a component with its environment is carried out via

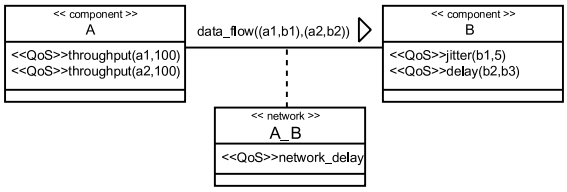


Figure 1: A Configuration of Components in UML Class Diagram

input and *output* signals. As a result, interfaces of a component specify signals that the component receives or emits.

Each component is associated with a number of input and output signals. In this paper, signals are denoted by x, y and z. Time of occurrence of a signal is denoted via a non-negative sequence of rational numbers. For example, the time of occurrence of a signal x is denoted with $x_1, x_2, \ldots$ representing time of first, second, ... occurrence of x.

Timeliness QoS expressions [9] such as jitter, throughput and latency can be expressed via first-order logic formulas on the set of time of occurrence of signals.

Throughput of at least (most) K within the time period T, for signal x can be written as the first order formula $\forall i \in \mathbb{N}: \quad x_{i+K-1} - x_i \leq T$ ($\forall i \in \mathbb{N}: \quad x_{i+K-1} - x_i \geq T$), respectively.

Notice, paper [5] refers to the above QoS constraint as Non-Anchored throughput.

Jitter , also called Non-Anchored jitter , of a signal x can be defined by the expression $\forall i \in \mathbb{N}: \quad T - m \leq x_{i+1} - x_i \leq T + M, \quad$ where T is the period of the jitter and m, M are constant rational numbers.

Latency of at most T unit of time between two signals x and y as $\forall i \in \mathbb{N}: \quad 0 < x_i - y_{Ki+K'} \leq T$. A special case of the above definition (for $K = 1$ and $K' = 0$) is the well-known definition of latency $\forall i \in \mathbb{N}: \quad 0 < x_i - y_i \leq T$ that applies to the time difference of the i-th occurrence of x and y.

3 UML/OCL based design of real-time systems

A real-time system can be designed as a set of components where signal communication links exist among pairs of components. We can describe such components in a UML class diagram in Fig. 1.

Type Component can be specified by Stereotyping "component," by which user can easily extend UML specification. A signal communication can be specified with Association.

Each of components has provided QoS, which can be represented via OCL annotation. Each of network links (which has association class with stereotype "network") also has network properties represented via OCL annotation. The network properties is the same as timeliness QoS. Attribute region of each class, includes special variables for QoS with "QoS" stereotype (in Fig. 1).

The following is the syntax of the variables.

Throughput Variable := "throughput(" signal "," period ")" ;
Jitter Variable := "jitter(" signal "," period ")" ;
Latency Variable := "delay(" output "," input ")" ;

A class with "component" stereotype has three categories of timeliness QoS (Jitter, throughput and latency), while a class with "network" has two categories of timeliness QoS (Jitter and latency).

The OCL description is given as follows[9].

QoS description := "context" className invariant* ;
invariant := "inv: self." constraint ;
constraint := variable op constant;
variable := Throughput Variable | Jitter Variable | Latency Variable
op := ">" | "<" | "≥" | "≤" ;

For example, the following are examples for Fig. 1, where – means a comment line.

context A
inv: self.throughput(a1,100) ≥ 20
– signal a1 is emitted at least 20 times in the period 100 units of time
inv: self.throughput(a2,100) ≤ 10
– signal a2 is emitted at most 10 times in the period 100 units of time

context B
inv: self.jitter(b1, 5) < 1
– signal b1 has jitter 1 with period 5 units of time
inv: self.delay(b2,b3) < 5
– latency between receiving signal b2 and sending signal b3 is less than 5 units of times

context A_B
inv: delay ≤ 100
– latency (network delay) between component A and component B is less than 100 units of time

4 The Verification Method

The verification consists of two steps; First Step and Second Step.

If some of the components are not simple enough, then repeat the process again from First Step on each of the components.

The following is the abstract level of steps of the proposed method.

input:

- system required timeliness QoS represented in OCL;
- component level provided timeliness QoS represented in OCL; and
- network configuration represented in UML/OCL class diagram.

output:

- component level behavioral specification represented in UML/OCL state-chart which satisfies required timeliness QoS under the configuration;
- or failure.

1. First step
   (a) We generate test automaton form the required timeliness QoS.
   (b) We generate abstract QoS automaton form each of the provided timeliness QoS.
   (c) We generate configuration automaton form the network configuration.
   (d) We check the consistency from parallel composition of the above automata.
   (e) If the result is deadlock then return failure. We have to reconfigure the requirement or provided conditions.
   (f) If the result is not deadlock then go to Second Step.
2. Second Step
   (a) If the component is not small enough to represent simple state-chart, then refine the component by;
      i. renaming provided QoS of the component to required QoS;
      ii. design sub components and provided QoS of each of them;
      iii. design network configuration; and
      iv. we repeat the First Step until the component is enough to small.
   (b) If the component is small enough to represent simple state-chart, then we describe state chart of the component.
   (c) We translate a test automaton from the provided timeliness QoS.
   (d) We design network of timed automata from the state-chart.
   (e) We check the consistency from parallel composition of the above automata.
   (f) If the result is deadlock then return failure. We have to reconfigure state-chart.
   (g) If the result is not deadlock then return success.

4.1 The First Step

Verification inputs are the following.

- Required QoS;
- a set of components with provided QoS; and
- a configuration automaton which represents network properties.

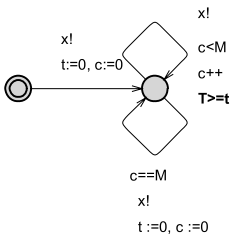


Figure 2: Abstract QoS automaton for Anchored Throughput

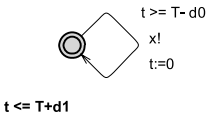


Figure 3: Abstract QoS automaton for Non-Anchored Jitter

The output is whether a given required QoS is satisfied under a given set of components with provided QoS and a given set of network links with network properties.

In usual methods, designer models behavior of each component in a network of timed automata and for the whole network of timed automata. Then the designer performs model checking, which often results in state explosion.

Here, we give a new method, in which timed automata (We call each of them an **abstract QoS automaton**) is derived automatically from the provided QoS.

The important point is that derived automata are so small that state-explosion is avoided.

Here, we give a translate rule for each provided QoS.

4.1.1 Throughput

For a component with anchored throughput with parameter max:M, min:m and period:T, we give an abstract QoS automaton in Fig. 2.

4.1.2 Jitter

For a component with non-anchored jitter with parameter max:M, min:m, we give an abstract QoS automaton in Fig. 3.

4.1.3 Latency

For a component with latency with parameter max:M, min:m, we give an abstract QoS automaton in Fig. 4.

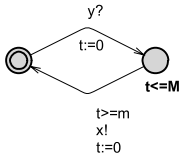


Figure 4: Abstract QoS automaton for Latency

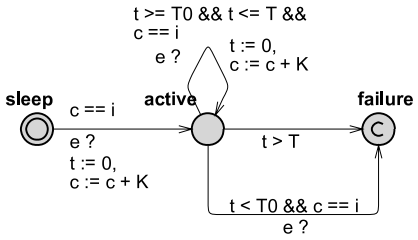


Figure 5: Test Automaton for throughput

Unfortunately, the automaton does not accept input y until it emits output x. To avoid the problem, a set of the same automata is needed. The number of automaton decided from throughput property of the components.

4.1.4 Configuration Automaton

A configuration automaton models interfaces among components. As each component has several inputs and outputs, such an I/O is represented as a channel in the configuration automaton. Each channel synchronizes with some I/O of some components with provided QoS (abstract QoS automaton). Abstract QoS automata and the configuration automaton communicate each other as described above.

4.1.5 Test Automaton

Here, we give each test automaton for throughput, jitter and latency[10].

Throughput For a non-anchored throughput of which a signal e occurs at least k times in a period T, a network of test automata consisting of k processes of timed automata in Fig.5 observes the throughput. In the network of test timed automata, the variables c is shared among automata globally. Each of timed automata is activated by turns along the value of variable K.

Jitter Figure 6 shows a test automaton for anchored jitter. It observes whether a signal e occurs periodically in the period $[nT - delta_0, nT + delta_1]$, where $n = 1, 2, 3, \ldots$.

Latency Figure 7 provides a component of test automata for latency between a signal x and y.

T/D processes of component in Fig. 7 observes latency at most D unit of times of the latency.

4.1.6 Verification

The behavior of such media with timeliness property also is modeled in a network of timed automata, we call such an automaton a configuration automaton.

Parallel composition of an abstract automaton for every component and the configuration automaton and test automaton for specified timeliness QoS decides whether the whole system satisfies the specified timeliness QoS.

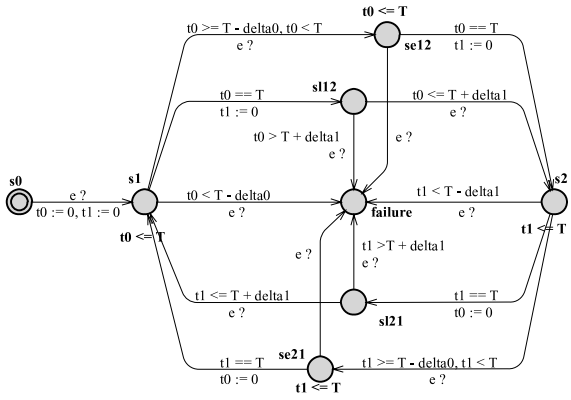


Figure 6: Test Automaton for jitter

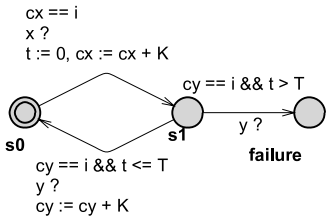


Figure 7: Test Automaton for latency

For the detail of process of the verification, refer [10].

4.1.7 Category Based Model Checking

Verification is performed for every timeliness QoS category (latency jitter and throughput). The idea and approach is very simple. When we want to check only latency as the required QoS, we build an abstract automaton for provided QoS of latency only.

The divided and conquer approach, reduces the size of states.

4.2 The Second Step

For each of components, Second step has the following two cases depending on the component's abstraction

- We repeat First step to the given component recursively.
- We design detail behavior of the component and verify whether provided QoS is ensured by the design.

If the size of the given component is large and designer has to design the given component from more detail components, then repeats First step.

Hereafter, we describe the later case.

At Second step (of the later case), verification is independently performed for each component.

Before Second step, the designer has to give detailed behavior of each component. Such behavior is given in UML state-chart. In order to give time constraints on events, the state-chart has clocks.

Verification inputs are the following.

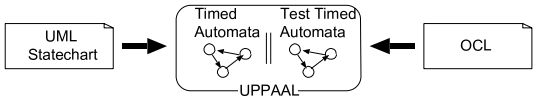


Figure 8: Verification on UPPAAL based on Test Automata

- component's behavior given in UML state-chart with clocks; and
- component's provided QoS.

The output is whether provided QoS is satisfied under a given UML state-chart with clocks.

The verification is performed based on test automaton. We have to translate a UML state-chart with clocks to a network of timed automata.

A state-chart can represent hierarchical architectures; while a network of timed automata is a simple flat structure model. In general, hierarchical structure can be flatten, but such translation increases the number of states. There are several translations, and this paper adapts the one in [4]. The translation itself is an algorithm to translate Hierarchical Timed Automata (HTA) to a network of timed automata used by UPPAAL[11].

Thus, we have to translate state-charts to HTA. Fortunately, syntax and semantics of state-chart and HTA are both similar, the translation is simple.

We add the following constraints on the state-chart.

- the state-char diagram has clocks; and
- arcs in the stat-chart has clock constraints in a form of the one same as Timed automata in UPPAAL.

We also use test automata to check timeliness QoS. Test automata for jitter, latency and throughput are given in 4.1.5.

The verification can be performed with UPPAAL. Thanks to test automata, we just check deadlock property for each QoS. Logical expression for deadlock property in UPPAAL is "A[] not deadlock."

5 Experiment

The proposed method is applied to an example.

5.1 The example

Media Server is an application delivering video stream and audio stream to Digital Television and Audio System[6], [12]. Each of output devices required timeliness QoS (throughput). Figure 9 shows the class diagram of the application, which consists of twelve components.

In order to compare the proposed method to the old method, which uses LP solver to First Step, we merge the twelve components to three components (Server 3 components , Audio client 4 components, and Video clients 5 components).

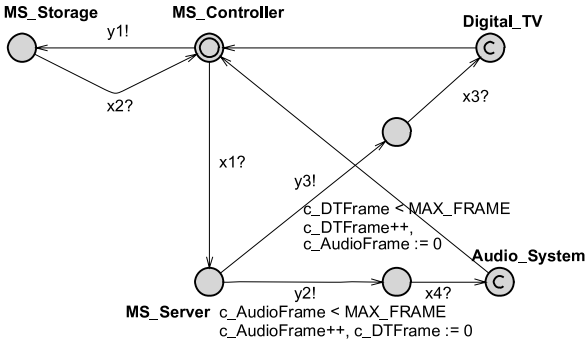


Figure 10: The configuration automaton

5.2 First Step

The following is the provided QoS.

- Throughput of Component MS–Server is equal or greater than 100 frames/s.
- Processing latency of Component MS-Storage is equal or less than 5ms.
- Network latency between MS–Server and Digital–TV is equal or less than 100ms.
- Network latency between MS–Server and Audio–System is equal or less than 150ms.

For these provided QoS, and a configuration automaton derived from the UML class diagram, and Required QoS for the whole system, we apply the verification along with First Step.

Here, we performed verification under the condition that the throughput of Digital Display (DT Display) is over 30 frames/sec as Required QoS.

Figure 10 shows the Configuration automaton for the experiment. The Configuration automaton in Fig.10 represents connection among the components. It uses channels to communicate abstract QoS automata providing the provided QoS mentioned above. For example channel x is used for communication to an abstract QoS automaton with throughput 100 frames/sec at a transition between MS–Controller and MS–Server. In order to avoid unfairness that frame communication occurs only between MS–Server and Digital–TV (or only between MS–Server and Audio–System), we use a parameter MAX_FRAME, which is used in a condition that the maximum successive occurrence of signals between the same components. We use a condition $MAX\_FRAME = 1$ for the experiment.

Figure 11 shows the test automaton for throughput as Required QoS. The test automaton observes throughput of 3 frames per 100ms. As required QoS, the required value of throughput is 30 frame/sec, We have to need 30 processes of throughput test automata to observe the throughput exactly.

At the First Step, we have performed verification experiments for two configurations: (1) an abstract QoS which outputs ten frames per 100 msec, and (2) an abstract QoS which

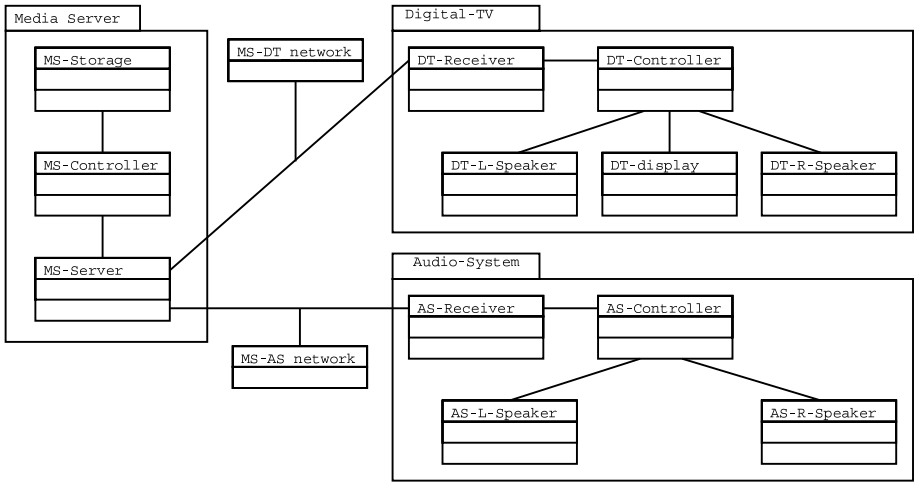


Figure 9: Class Diagram of Media Server

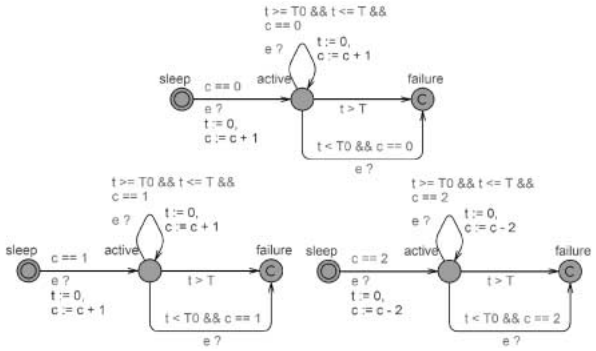


Figure 11: The network of test automata for the given required QoS

Table 1: The result (1) of first step

# of P	result	CPU time	Used memories
3	not valid	0.3 ms	23.9MB
6	not valid	1.4 ms	24.4MB
9	valid	28 ms	24.8MB
12	valid	50.6 ms	25.9MB
15	valid	83.6 ms	26.8MB
30	valid	480 ms	34.4MB

outputs 30 frames per 300 msec, respectively, are used for abstraction of provided QoS for MS.Server. Each of two experiments is performed with several numbers of test automata: 3, 6, 9, 12, 15 and 30. We have obtained CPU times and sizes of memory consumed. The experiments are performed in the following environment: CPU is Intel Core 2 Duo 2.33GHz, OS is Windows Vista Business and M.M. is 2GB. We used UPPAAL4.1.0 as a model checker. Table 1 and Table 2 show the results of (1) and (2), respectively. The column of "# of P" shows the number of processes (the number of test automata).

Table 2: The result (2) of first step

# of P	result	CPU time	Used memories
3	not valid	0.6 ms	23.9MB
6	not valid	0.7 ms	24.4MB
9	not valid	1.2 ms	24.7MB
12	not valid	1.6 ms	25.0MB
15	not valid	2.1 ms	25.1MB
30	valid	957 ms	40.4MB

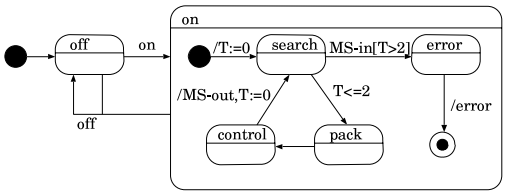


Figure 12: A UML Statechart Diagram of Component MS-Storage

In the previous experiment, we have performed First Step with Linear Programming solver. In the experiment, we have performed it in 78 ms (although it has been performed in different environment).

5.3 The Second Step

After First Step, we design inner behavior of each component.

In the example, recursive application of First Step is not performed, because each component is small enough. Behavioral specification is described in UML state-chart. The design must meet the provided QoS. Figure 12 shows behavioral specification of Component MS-Storage.

In order to verify timeliness QoS for each component, Statechart must to be translated into a network of timed automata

and also timeliness QoS is converted into test automata.

Figure 13 depicts translated result of *on* part in Fig.12.

The translating times are summarized as follows.

Translation time	: 1153 ms
The number of states(before)	: 89
The number of states(after)	: 179

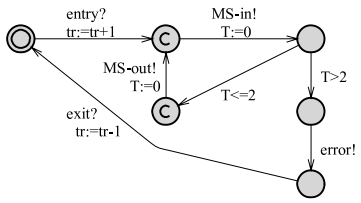


Figure 13: An UPPAAL Timed Automaton of Component MS-Storage

For every component and for every timeliness QoS, verification is performed. The total CPU time is about one seconds.

We found that for every component, the verification is performed within a few seconds with UPPAAL, without state explosion. Also we found that there is no deadlock.

5.4 Discussion

Table 1 shows that when we perform the experiment with nine test automata, it outputs the correct result. The result of Tab.2, however, shows that we cannot obtain the correct result until the number of test automata increases to 30. When we perform the experiment with 30 test automata, the CPU time increases exponentially. Therefore, we can conclude that there is a trade-off between degree of precision and CPU times. As shown in both tables, the CPU times of the experiments with 30 test automata are too large. Thus, as we consider the trade-off, in this experiment, the trade-off point is at which the number of process is 15.

Though we cannot exhibit that our proposed method is better than that of linear programming based method with respect to the CPU time, the performance of the proposed method is within useful reasonable time. The linear programming method has many constraints on configuration, while the new proposed method is flexible and is able to apply recursively along with component hierarchy, which are the advantage of the proposed method.

Our proposed method has more acceptable inputs than the former method. The difference between this and that of general class is very small. It is although not faster than the former method, it is more flexible than the former method.

6 Conclusions

This paper proposed a stepwise verification method for design of real-time system with UML/OCL focusing on timeliness QoS aspects. The method uses abstract QoS timed automata in order to reduce the possibility of state explosion.

The method can be applied to a design with complex connection of components.

Future works include simultaneous verification of several kinds of timeliness QoS, and utilization of feedback information such as verification counter-examples.

Acknowledgement

This work is being conducted as a part of Stage Project, the Development of Next Generation IT Infrastructure, supported by Ministry of Education, Culture, Sports, Science and Technology.

REFERENCES

[1] R. Staehli, F. Eliassen, J. Aagedal and G. Blair "Quality of Service Semantics for Component-Based Systems," 2nd Int'l Workshop on Reflective and Adaptive Middleware Systems, pp.153-157, 2003

[2] R. Alur and D.L. Dill: "A Theory for Timed Automata," In Theoretical Computer Science 125 pp.183-235, 1994.

[3] L. Ageto, P. Bouyer, A. Burgueño and K. G. Larsen: "The Power of Reachability Testing for Timed Automata," LNCS, Vol.1530, pp.245-256, 1998.

[4] A. David and M. O. Möller: "From HUPPAAL to UPPAAL: Translation from Hierarchical Timed Automata to Flat Timed Automata," BRICS Technical Report Series, RS-01-11, 2001.

[5] H. Bowman, G. Faconti and M. Massink: "Specification and verification of media constraints using UPPAAL," In Proceedings of Design, Specification and Verification of Interactive Systems'98, pp.261-277 Springer, 1998.

[6] K. Havelund, A. Skou, K. G. Larsen and K. Lund: "Formal Modelling and Analysis of an Audio/Video Protocol: An Industrial Case Study Using UPPAAL," In Proceedings of the 18th IEEE Real-Time Systems Symposium, pp.2-13, 1997.

[7] Object Management Group: Unified Modeing Language Specification version 2.1, available at http://www.omg.org/.

[8] B. Bordbar, J. Derrick and A. G. Waters: "A UML approach to the design of open distributed systems," In Chris George and Huaikou Miao, editors, Formal Methods and Software Engineering, LNCS, Vol.2495, pp.561-572, 2002.

[9] UML Profile for Modeling Quality of Service and Fault Tolerance Characteristics and Mechanisms, Request for Proposal, available at `http://www.omg.org`

[10] B. Bordbar and K. Okano, "Verification of Timeliness QoS Properties in Multimedia Systems," 5th International Conference on Formal Engineering Methods (ICFEM '03), LNCS 2885, pp.523-540, 2003

[11] Johan Bengtsson, Kim G. Larsen, Fredrik Larsson, Paul Pettersson, and Wang Yi: "UPPAAL . a Tool Suite for Automatic Verification of Real-Time Systems," LNCS, Vol.1066, pp.232-243, 1995.

[12] D. Akehurst, J. Derrick, and A. G. Waters: "Design and Verification of Distributed Multi-media Systems," LNCS, Vol.2884, pp.276-292, 2003.

[13] The official OMG SysML site, http://www.omgsysml.org/
[14] J. Carlson, J. Håkansson and P. Pettersson: "SaveCCM: An Analysable Component Model for Real-Time Systems," Proc. of FACS'05, Electronic Notes in Theoretical Computer Science, Vol.160, pp.127-140, 2005.
[15] E. Nagai, A. Makidera, K. Okano, K. Taniguchi: "A Method to Develop Distributed Real–Time Applications Based on UML/OCL (in Japanese)," In the IEICE Transactions, Vol.J89-D No.4, pp683-692, 2006.

A Proposal for a New System Structuring Method Supplementing Technological Development of Semiconductors

Kunihiro Yamada [1], Kouji Yoshida [2], Masanori Kojima [3], and Tadanori Mizuno [4]
Tokai University
2-2-12, Takanawa, Minatoku, Tokyo, 108-0074, Japan yamadaku@tokai.ac.jp
Shonan Institute of Technology
1-1-25, Tujido Nishikaigan Fujisawa Kanagawa 251-8511,Japan yoshidak@info,shonan-it.ac.jp
Osaka Institute of Technology kojima@is.oit.ac.jp
Shizuoka University mizuno@inf.shizuoka.ac.jp

Abstract: All of the system and semiconductor, such as electronic equipment, played and fulfilled their respective roles as a system and its parts, and were in mutual prosperity in a half past century. However, the performance of semiconductors exceeded system requirements and semiconductors with reasonable cost have been provided in some systems since 1990s. After this time of period, both hardware and software have presented more complicated situation than required, and this situation caused a crucial quality problem of a developed system after it was released in the market or longer development man-months period than traditional presumptions[1]. The major cause of above problems is supposed that the design was not performed as a system and that the functional differentiation of the roles between hardware and software was not conducted adequately.

We will present two systems under study and propose a new system structuring method. This system structuring method is to realize system performance by cooperation between hardware and software, and enables to produce system structure that minimizes the total necessary cost for development, production and application of the system.

***Keywords*:** System, Semiconductor, System Structure, Hardware, Software

1 INTRODUCTION

Many kinds of equipment were born thanks to the existence of semiconductor. Therefore, the role of semiconductor contributing to modern electronic equipment is very important. Many kinds of semiconductor-applied products such as communications equipment, home electric appliances, computers, have been widely developed thanks to semiconductors but also there are many systems that had existed before semiconductor was invented. Many questions may be answered when comparing these equipment, devices, and tools prior to the use of semiconductors with ones after the use of them. What we expect it to do is reminding people of the original vigorous features of equipment, device and tool. Magnificence of mechanism should also be re-acknowledged. The system development has to be started including all of these features. This will be effective both in global energy saving and in business activity.

In 1990s, semiconductors with sufficient capability and with reasonable cost were provided for engineers who were getting familiar with software and hardware design. As a result, the development in the extension of tradition has been backed-up and deployed by the aid of an enhanced CAD. Even if system structure were not the best one or wrong one, CAD would move the system. The real designer does not exist and an operator of CAD masquerades the designers' position. Then the semiconductor that is improving its capability is departed from the region of real master. As a result, both software and hardware engineers who forgot the system have lost their coordination and done nothing but grow in their own way, and the situation has continued to get more complicated than required.

The object of the system structuring method presented here is to work out system structure that minimizes the total cost of system development, production and application. Even if performance is the same, system structures may differ when manufacturing and system application conditions are different. As examples, two system cases currently under reviewing are presented. The first system is characterized by many series development, and the second system is characterized by requiring a lot of adjustment in actual system installation and operation. For them, the system that can minimize the total cost has been obtained. This paper shows that the system capability can be realized by cooperation between hardware and software.

Namely, assuming that semiconductors will show more progress in future, we describe wafer processing and assembly technologies that play leading roles in semiconductor progressing. Also, we describe the status of mutual evolution of semiconductors and electronic equipment applied to computers and videocassette recorders as an example since the relationship between semiconductors and system will be more important in future.

2 DEVELOPMENT OF SYSTEM AND SEMICONDUCTOR

2.1 Mutual Development

There are three important items in the development of electronic equipment, the first one is the specification of market requirements, the second one is product

characteristics such as capability, performance, and quality, and the third one is the cost related to development and production. In connection with them, semiconductor has contributed to electronic equipment bit by bit since mid 1960s. Also, various standardized semiconductors have been developed preoccupying the market toward general-purpose application in order to use semiconductors in various electronic equipment.

The designers of electronic equipment started to join semiconductor design in 1970s and typical examples are software design for microprocessor and logical design of gate array. By the above trend, application of semiconductor to electronic equipment became easier. The successful application of a semiconductor to some electronic equipment realizes technology distribution to other electronic equipment via standardized semiconductor goods. Thus, while electronic equipment and semiconductor have made repeated efforts of technological development in each area, they made the outcome cross-fertilized each other to progress.

2.2 Development of Semiconductor Technology

The finding of detection mechanism by point contact method in 1874 was the origin of semiconductor [2]. Due to the unstable performance, however, diode or triode vacuum tubes, invented just after this finding, were used. A point contact type transistor was invented at Bell Telephone Labs in 1947 and a junction type transistor was invented in 1949 [3] . These epoch-making inventions realized vacuum tube capability in compact size, light weight and with low power consumption and brought out many tangible and intangible benefits such as significant reduction of time and cost required for appropriate system operation.

Generally, semiconductor technology means following four items: (1) manufacturing technology such as wafer processing, assembly and test; (2) engineers' idea and design technology such as various simulators and CAD; (3) system technology regarding specifications of semiconductor device; and (4) application technology of semiconductor device to electronic equipment. The technological progress is described in this paper by concrete examples regarding manufacturing technologies, wafer processing technology and assembly technology, which act as an accelerator of miniaturization.

(1) Wafer Processing Technology

The transistor has evolved into integrated circuit, and improvement of capability and incredible expansion were realized by the progress of semiconductor manufacturing technology, especially miniaturizing technology for wafer processing. The strong demand of cost reduction that makes the price of chip unit lower by advancing miniaturization and specific property of semiconductor in order to realize performance improvement such as speeding up and reduction of power consumption, has promoted the technology. However, it is said that the performance improvement by scaling is not expected anymore to overcome several issues simultaneously since trade-off relation exists when the gate length is below 10 nm after 2020 [4]. The miniaturization of wafer processing is summarized in **Fig. 1**, including degree of integration and operation speed of semiconductor component.

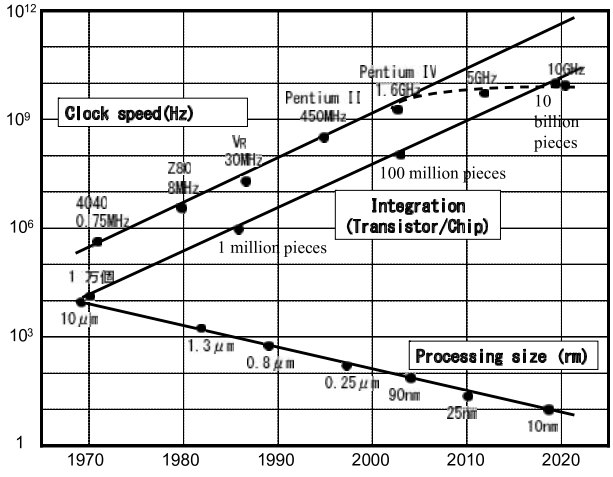


Fig.-1: Miniaturization of Wafer Process and Semiconductor Component/System

Miniaturization has been advanced by two digits, from 10 μm to 0.13 μm, and the number of elements (integration degree) has been increased by four digits, from 10 thousand to 100 million from 1970 to 2003. From now, it is expected that the miniaturization will be advanced by one more digit and the number of components will be increased by two digits to 10 billion by 2019. On the other hand, the clock frequency of microprocessor will be increased by four digits, from 0.75 MHz in 1971 to 10 GHz in 2019[5]. At the moment in 2008, there are microprocessors with 600 million transistors in one chip or multi-valued NAND flash memories of 2, 3 or 4 bits [6].

(2) Assembly Package

Accompanied with miniaturization, dramatic increase of production cost (e.g. mask cost of 100 nm being up to ¥100 million) and protracted development period are attracting attention, and SoC (System on Chip), called ultimate form, is no longer a realistic solution. SiP (System in Package) showed up with two chips contained around 1998 is effective in the field of cell-phone and up to seven chips are contained in one package now.

In 1971 when four bit MPU appeared, the gate length was more than 5 μm and the number of pins on a package was two digits. However, accompanied by the increasing range of functions of one chip realized by the later miniaturization of wafer processing, the number of assembly pad on a chip

and the number of pins on a package have been increased. Future perspective is shown in **Table 1** [7]. Requirements for IC package are an increased number of pins corresponding to higher integration, resistance to lower temperatures corresponding to higher heat radiation, optimization of electrical characteristics corresponding to higher speed, higher density corresponding to miniaturization and multifunction, etc. Also, from the viewpoint of higher package density of chips, chip integration technique, through-hole connection of a chip, COC (Chip on Chip), flip chip connection by bump between chips, WPP (Wafer Process Packaging), rewiring on a chip, etc., are also under consideration.

Table 1: Increasing of the Number of LSI Pins Accompanied by Miniaturization

Year (Gate Length of MPU, nm)		1971 (5000)	2003 (65)	2010 (25)	2018 (10)
Number of Pads (Max.) [2]		100	3072	3840	4544
Number of Pins (Max.) [2]	CPU	40～100	1452	2762	5426
Actual Number of Components/ Pins	CPU	4004/ 42	Emotion Engine/ 540	? 900	? 1800
	Logic	100	2400	4009	8450
	Logic	14～100	2034 (Renesas Flip Chip Grid Array)	3500?	7000?

* Model Name of CPU/MPU Trend of Number of Pins

Number of Bits	4	32	64	128
Model Name	4004	80386	R4900	Emotion Engine
Number of Pins	40	132	304	504

(3) Type of Semiconductor

Figure 2 shows a process of a system consisting of 12 LSIs integrated into one LSI.

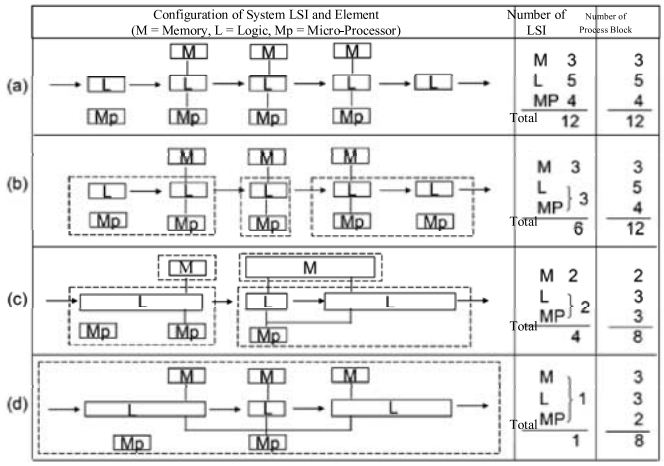


Fig. 2: Relationship between Number of LSI Reduction at System LSI or SoC Progress and Number of Processing Block

This is realized by improving processing method, miniaturization of wafer process, and faster processing. From (a) of Fig. 4, wafer processing for 0.5 μm, 0.25 μm, 0.13 μm, and 0.09 μm are shown respectively.

Table 2 shows an example of consideration on the sequence how various semiconductor devices are integrated in a system from the viewpoint of SoC and SiP. Among semiconductor chips integrated into customers' chip set circuit design, DRAM with different production process, for example, and analog device are sometimes mixed, which means it would also be involved with process technology development when heading toward SoC.

Table 2: Layering of Semiconductor Device Toward System LSI

	Specific Semicon ductor	LSI	V~ULSI		System LSI		
					SoC	SIP	Remarks
Semicon ductor Device	FET	MOS IC	Memory	DRAM, SRAM SDRAM, Flash	• MCU + ASSP (ASIC for TV) • CPU + USB + Flash + SDRAM • DSP Core + Memory for MCU, etc. (A part of cell-phone capability) • CPU + USB + SDRAM + JPEG/ MPEG Performance LSI + (DSC System) • Audio Engine + USB + CPU + (HDD/Network Towarding Audio)	• Driver + MOS FET • Stacked Type CSP Memory (SRAM, Flash, etc.) • CMOS SP Logic + BiCMOS ASSP • SoC + Analog LSI + Large Capacity Memory • Application Processor + SDRAM + … (Cell-phone)	SoC → SIP → SoC → SiP I • Being Miniaturization • Being High Integration • Being High Performance • Reduction of Number of Chips • Cost Lowering Depending on the above demands, being SoC and SiP are repeated.
			Micro	CPU, MPU, MCU, DSP			
			Logic	MSIG, FPLD GP/SP Logic			
	Bip Tr	Bip IC	Digital Bip.	Multipurpose Logic ASSP			
			Analog Bip				
Major Applicati on Products	• Transistor Radio	• TrTV • Calculator • Clock	• PC • Printer • ODD • Game Machine	• VTR • CTV • Operation Amplifier • A/D Converter	• Cell-phone • DVD • DSC • Household Electrical Appliance Device for Digital Information • Information - Communication Equipment	• Cell-phone • DSC • CIS • PDA • Household Electrical Appliance Device for Digital Information • Information - Communication Equipment	
Remarks	• Wrap up to LSI of specific semiconductors (Integration) is the first step toward SoC in future.		• MPU consisting of CPU core and peripheral logic is also a kind of being SoC. • Performance of a part of application product is realized by LSI.		• Improvement of Quality, Performance and Capability by Reducing Number of Chips • A Part of Performance in Application Product Fulfilled by SoC	• Being SoC is difficult for processing technology. • Bringing the demands of high performance, low cost, flexibility, etc., into real by SoC is difficult.	

2.3 Progress of Electronic Equipment (System)

Even though assuming that we did not have any semiconductors, we could have computers, videocassette recorders, cars, telephones, airplanes, etc. We could own a car, take air plane, which makes no difference in our life and work leading no inconvenience. On the contrary, there may be no privately owned PC, videocassette recorder, cell-phone as we have today, but other machines and mechanism might well refined to be practical other way around. Though semiconductors are necessary for human beings to travel in space, they are considered to be not sufficient for it.

Relationship between semiconductor and technology that supports the improvement of electronic equipment is described below by taking computer and videocassette recorder as examples.

(1) Computer

A computer has started its role from merely calculating and has enhanced its capability to the level of control, communication, data processing, computer graphic, network and even artificial intelligence. The first computer in the world, i.e. the original form of current computer, was EDSAC (Electronic Delay Storage Automatic Calculator) developed by M. V. Wilkes, et al, in 1949 [8]. Several concepts of fundamental method for establishment of current computer and the trend of realizing technology and production quantity are shown in **Fig. 3**. The fundamental configuration of computer with most important concepts was completed within only 10 years after the successful development of EDSAC in 1949.

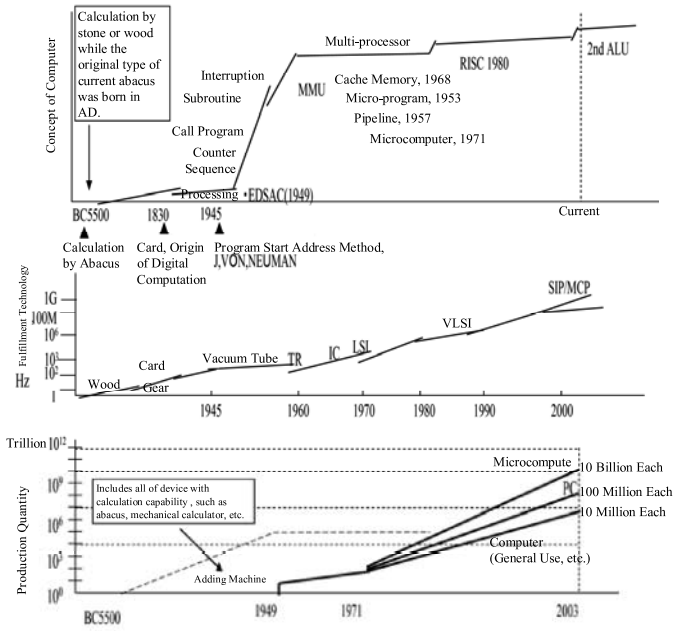


Fig. 3: Enabling Technology for Computer and Production Quantity

The semiconductor has played a large role in the subsequent realizing technology of computer evolution and the semiconductor itself has improved as well. The success is typically represented by the production quantity that shows 10 million of general-purpose computer, 100 million of PC, 10 billion of microprocessor. It can be considered that even though 10 million of general-purpose computers may exist, 100 million of PC and 10 billion of microprocessor would not exist if it were not for the semiconductor.

There are two factors for the success of computer. One is the innovation of program storage method to realize computer and the other is the semiconductor. The program storage method is a function appropriate to be called as general processor, and is able to configure wide variety of algorithms and processing systems.

(2) Video Recorder

Analog circuits of Audio-Visual equipment using vacuum tubes have been replaced by transistorized equipment since 1965 with less space and decreasing failure rate. The current home VTR was developed from 1974 to 75 and the mechanical tape traveling was control by microprocessor. A part of analog circuit was digitized since 1990s to keep stability of the circuit characteristics and was applied to the simplification of adjustment in production lines.

Currently, signal itself is treated in digital form such as DVD, DTV and CD and data compression/expansion technology expands new possibility in communication and recording. It is considered that elements number in 10 nm wafer will be 100 times in 2019, and that more than 200G bits (25 GB) memory in one chip is possible. The semiconductor memory will be practical in video recording, 2.5 hours (MPEG-2) for HDTV, 9.2 hours (MPEG-2) for SDTV and 27.5 hours for H264. In other words, it is considered that the video recorder will also advance from tape, disk and to semiconductor led by video media in long years.

Performance expansion of VTR and historical trend of semiconductor chip are shown in **Fig. 4** [9]. Two movements, one was reducing number of LSI by chip integration and the other was increasing number of LSI to expand performance, were in progress simultaneously. The performance enhancement saturated at certain time, and the chip integration was accelerated. Then the LSI was finally integrated into two or even one.

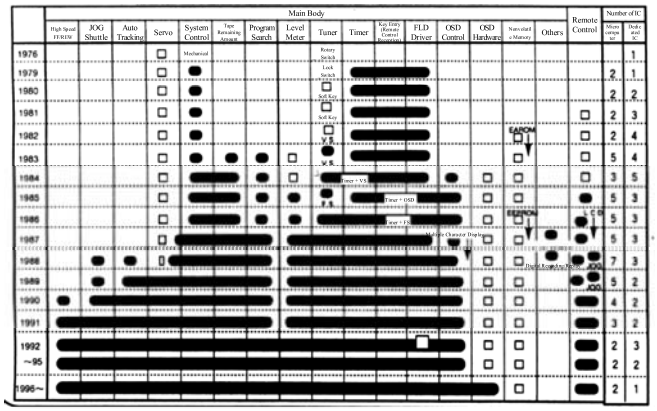


Fig. 4: LSI Consolidation at VTR

Next, processing method of Servo Motor is shown in **Fig. 5** [10]. This is also the example of chip integration as mentioned above. Analog servo for motor drive was bipolar-process, but the integration was made in the same chip since the digital servo is the same CMOS process as microprocessor. After that, the dedicated circuit for digital servo became one digit lower in logic circuit when moved to

software servo and major remaining components were only a simple counter and a PWM circuit. The shifting of servo circuit from analog to digital or digital to analog is based on the technology improvement. The former is by the increasing number of components on one chip and the latter is by the speeding up of processing speed of microprocessor. Both are thanks to the performance improvement brought by miniaturization of wafer process.

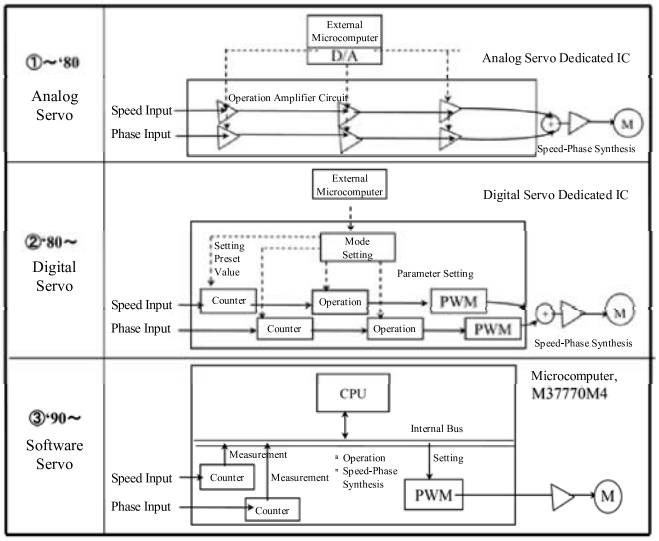


Fig. 5: Evolution of Motor Servo Processing Method

The features of software servo are smaller number of external circuits compared with analog servo, and also small number of process for design improvement and shipping test after manufacturing. Also, several learning functions can be installed by motor revolution sensing since it is processed by software. They are voltage fluctuation, aging fluctuation, motor characteristics, etc.

3 NEW SYSTEM STRUCTURING METHOD

Since 1990s, among several systems such as electronic device, quality problems of developed system causing serious malfunction after they were released in the market or the problem of more development man-months than had expected was experienced. It was considered that the major cause was because design had not been performed as a system.

This paper proposes a new system structuring method as the initial point of system development. This system structuring method is to produce system structure minimizing the total cost of system when it is developed and manufactured. Profit produced relative to the cost, social contribution, business capability, etc., are not evaluated here as they are already considered sufficiently. It is supposed that "system structuring for minimizing total cost" is applied to partial modification of the system software and hardware previously developed or already in the market in most of the cases. However, we will describe not only such cases but also two other examples of situation changing design of system, hardware and software partly or entirely.

The first one is characterized by series development for digital TV or DVD including compression/expansion of image and audio data. This case is dealt with a microprocessor with the 2nd ALU. The second case is a system of narrow area network such as targeted for home use, which is characterized by requiring individual adjustment for each installation and operation of the system. This case is dealt with two communication methods. In addition, we will explain that total cost can be reduced for each case.

3.1 Microprocessor with the 2nd ALU

The broadcasting system such as TV is roughly classified into three, NTSC, PAL and SECAM for analog, and four, HD and SD of respective 50 Hz power system and 60 Hz power system for digital. As for the display system, there is one for PC in addition to above and is roughly classified into eight, VGA, SVGA, XGA, SXGA, UXGA, WVGA, WXGA and WUXGA. By taking up DVD replay as an example to consider the sort of hardware corresponding to compression, expansion and display system, 15 sorts of hardware are required to be developed in order to output appropriate signals for all above TV receivers or PC monitors. If interlaced or progressive scanning is used differently, much more development is required.

Nevertheless, the development load is reduced by the cooperation of hardware and software. That is, the 15 sorts of compression and expansion as well as display function are executed by the program control of the dedicated microprocessor named 2nd ALU which is independent from the main processor bus. With respect to the system development including 15 sorts of hardware, a trial calculation has been performed showing that one kind of hardware dealing with 15 different kinds of programs which largely reduces development man-months (54 man-months, 23% of 234 man-months).

The development of hardware and software is easier by the miniaturization technology of semiconductor and development support tool of CAD compared to 10 years ago. Choosing which hardware should be conversed to software (program) or vice versa affects the basics of new system structuring. The method of MPEG compression and expansion executed by the program control of the 2nd ALU, which is independent from the main processor bus, is the conversion of hardware to software. It is important that development man-months can be reduced without making deeper layers or without increasing complexity of software in the main processor, and that processing is executed by a special processor independent from the main processor. In addition, there is a possibility that a dedicated processor is smaller than the random logic.

We have reviewed the processing method by the decentralization system independent from microprocessor by a bus switch, called 2nd ALU, for IDCT processing of MPEG-2 [11]. This was realized with small hardware of 1K gate logic, 2K RAM with 20 thousand transistors, and small software of 13 steps.

The concrete evaluation on the wafer area of the 2nd ALU is 17.4 % for ALU, and 2.3 % in total in case of microprocessor developed by 0.18 μm process. From this result, a decoder of MPEG-2 is expected to be constructed by 32-bit microprocessor with the 2nd ALU and SoC, and with DRAM mixed loading .

The structure of the 2nd ALU is shown in **Fig. 6**. The software exists in the small space of the 2nd ALU and the IDCT processing program in the MPEG-2 remains only 13 steps. This is the decentralized system giving independence of the 2nd ALU by bus switch in the 32-bit microprocessor. A 2K bite data memory, 4 sum of products and a control register are contained.

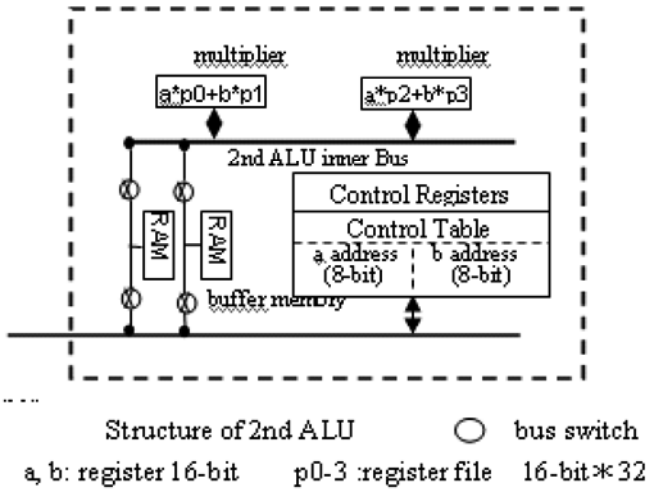


Fig. 6 Second ALU

We will evaluate the distribution of IDCT processing and data transfer time in case of 250 MHz clock bus speed at real time image processing. As for the IDCT processing time, the processing time for one buffer memory (8 blocks) is 13.3 μsec, the 2nd ALU is driven for 404 msec in one second. The transfer time between the microprocessor and the 2nd ALU requires 4.1 μsec with buffer memory of word configuration, 32 bit x 256, using read out after IDCT process and data setting. The transfer is executed 30,375 times within one second occupying the processor for 124 msec (**Fig. 7**).

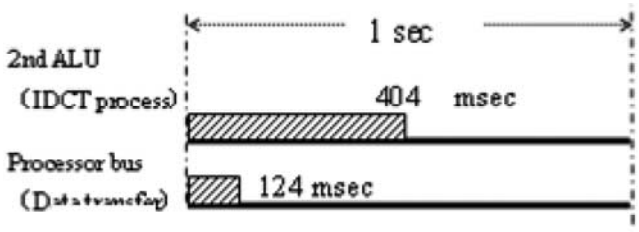


Fig. 7: Total process time for 1second

IDCT program at the 2nd ALU can be executed by 13 steps. Basically, the IDCT processing requires 128-step program. However, it was compressed down to 78 steps at Chen's Method [12]. The processing can be downsized to 13-step program by using the 2nd ALU. Further, by modifying the 4 sum of products in the 2nd ALU to 4 stanzas, the reduction to 4 steps is possible and the execution time in one second can be shortened as short as 124 msec.

3.2 Mutual Complementally Network System by Wire and Wireless Communications

As one of the effective system structuring method, the system capability is realized by cooperation between hardware and software. Namely, in this mutual complementary network system, hardware means two communications, wire and wireless, and software means mutual complementary performance. This improvement of performance is the trial not to complicate a system with increasing number of components by semiconductors miniaturization and this is not the improvement of existing communication performance but is realized by adding another communication capability as integration.

For more detail, 96% of communication performance at home was obtained by unifying two communication methods, 82% for wireless and 76% for wire at home, without any improvement in current communication environment for in-house network [10]. We named this communication system as "Mutual Complementally Network by Wire and Wireless Communications."

For home network, PLC (Power Line Communication) is promising media because it requires no additional wiring for communication. Recently, the improvement of communication quality of PLC is expected by employing spectrum diffusion technique with Chirp waveform at CEBus. The wireless communication is conducted according to IEEE 802.15.4, one of the standard of WPAN (Wireless Personal Communication Area Network) set by IEEE. The communication quality within certain number of packets decreases as the signal level of receiver decreases when the receiving antenna is distant from the output antenna. Also, when an obstacle is inserted between communication nodes, the interlayer of hierarchy becomes obstacle and the

deterioration of communication quality occurs, including the affection of noise, such as multi-path effect.

Configuration drawing (a) and its symbol (b) of house hold equipment which interfaces with network unit cell are shown in **Fig. 8**. The network unit cell shown in (a) consists of three sections: wireless communication, wire communication, and transmit/receiving data processing. The modeling of communication network was conducted both for single communication media and for double communication media through network cell. Simulation and experiment were conducted and the communication quality was confirmed to be improved.

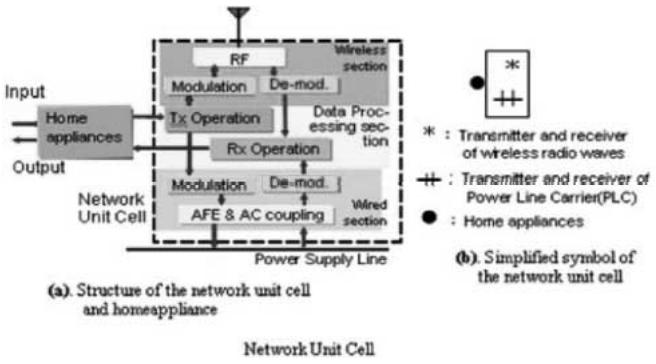


Fig. 8: Network Unit Cell

As the result shown in **Fig. 9**, communication performance for wireless of 82% and wire of 76% are increased up to 96% by the mutual compensation between two communication methods, and effectiveness of "Mutual Complementary Network by Wire and Wireless Communications" was confirmed[12]. It is also considered that this method is effective as one of the model of new system structuring.

This system is considered to be effective for the network that is used at home. It also seems applicable to the communication business area that requires higher reliability depending on its purpose[13].

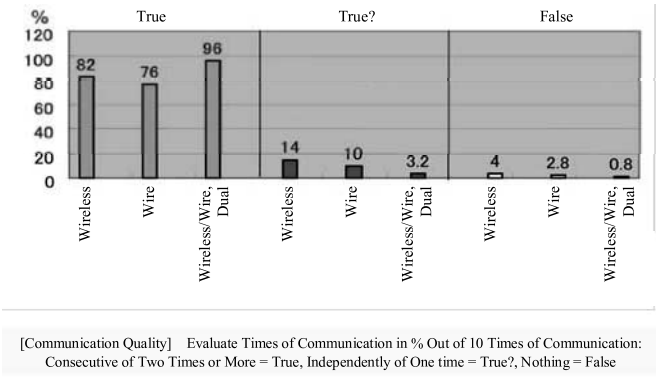


Fig. 9: Comparative Aspect of Communications Quality

4 CONCLUSIONS

We have proposed a new system structuring method as one of countermeasures and as the starting point of system development since both hardware and software are becoming too complex beyond necessity in these days. This system structuring method is to produce system structure minimizing the total cost of system when it is developed, manufactured and adapted to system, which decides the system structuring. New systems are not produced so frequently, but we have presented two examples of new systems. The first example is the case of digital TV or DVD including MPEG compression and expansion of image and audio data. A new hardware called 2nd ALU is proposed aiming at large reduction of development man-months. The calculation tying to identify its effectiveness showed reduction of 23%[14]. The second example is a system of narrow area network that is applied to home use, and is expected to be more standardized in future. Its individual adjustment at each location for installation and operation of the system will become a big amount of work. We have presented "Mutual Complementary Network by Wire and Wireless Communications" that improved performance considerably by the mutual compensation between wire and wireless in order to reduce individual adjustment. The above two cases make it possible to minimize the total cost of the system when it is developed, manufactured and adapted by obtaining an effective system structuring with cooperation between hardware and software.

It is required that system, hardware, and software be the best configuration based on the new system structuring method presented in this paper. This will also be the future subject of research, and from another point of view, it also means working out a mechanism and structure to reduce their capacity in system, hardware and software. This implies lower layer structuring as well as moving layer connection from plane connection to point connection as the structure with microprocessor having huge address space and some huge random logic (hardware) gradually decreases its huge random logic and the structure will consist of many microprocessors with small address space. The conversion of hardware to software (program) is not something stacking up software layers with huge address space but the system consisting of many microprocessors with small address space.

5 REFERRENCES

[1] NIKKEI ELECTRONICS Embeded Software 2006 ISBN4-8222-0249-2 p.22-24

[2]F.Braun "Uber die Stromleitung durch Schwefelmetalle. "Ann phys.Chem..153.556(1874)

[3] W.Shockley,"The Theory of p-n Junction in Semiconductor and p-n Junction Transistor,"Bell Syst. Tech.J..28.435(1949).

[4] Yosuke Mochizuki, "Evolution Scenario to the 10-nm Gate", October 2003 NIKKEI MICRODEVICES, pp.53-62.

[5] ITRS (International Technology Roadmap for Semiconductors) 2003 Edition, SIA.
[6] N. Shibata ,et al,"A 70nm 16Gb 16-Level-Cell NAND Flash Memory" 2007 Symposium on VLSI Circuits. Session 24.Embedded SRAMs, 18-5,2007. June 15
[7] Renesas Technology Corp. www.renesasu.com
[8] M.V. Wilkes and W. Renwick, "The EDSAC and Electronic Calculating Machine", Electronic Engineering, Vol.20, pp.385-391, 1949.
[9] Kunihiro Yamada, et al., "Applied Technology of One-chip Microcomputer and Software", Mitsubishi Electric Technical Review Vol.66, No.2, pp.229-235, 1992
[10] Kazuo Hayashi, Kunihiro Yamada, et al., "16-bit Microcomputer for VTR Software Servo", ditto, developed based on pp.205-213.
[11] K.Yamada,et.al."A New RISC Processor Architecture for MPEG-2 Decoding",IEEE Trans. On Cons.Elec., vol.48, No.1,pp.143-150,2002.2.
[12] W.A. Chen, et al., "A Fast Computational Algorithm for the Discrete Cosine Transforms", IEEE Trans. on Communications, Vol. COM-25, No.9, pp.1004-1011, 1977.
[13] Kunihiro Yamada, et al., "Dual Communication system using wired and wireless correspondence in a small space". KES2004,Springer LNAI 3214,pp. 898-904,2004
[14] Kunihiro.Yamada,et.al."Home-network of a Mutual Complement Communication System by wired Wireless " . KES2006,Springer LNAI4253,pp189-196 ,2006

Session 5: Groupware
(Chair Kenichi OKADA)

A Method of Collaborative Lecture which grasps the Level of Students' Comprehension and Its Evaluation Experiment

Takashi Dasai[1], Dilmurat Tilwaldi[2], and Hisao Koizumi[2]

[1] Department of Kansei Design, Hiroshima International University
[2] Graduate School of Science and Engineering, Tokyo Denki University

Abstract - The spread of internet has triggered a new trend of researches on the educational methods which incorporate the internet technologies. One of them is "Collaborative lecture", which aims at improving the effectiveness of the study by letting students work together in a collaborative manner so that the given theme is attained. Another one is "Distance learning", which conducts study via communication facilities between teachers and students who are mutually separated geographically. We believed that by incorporating the distance learning into collaborative lecture, we could improve the effectiveness of collaborative lecture. However, there was a tendency that collaborative lecture system was used independently from distance learning in the past, and there were very few work done in this scheme. In this paper, we will propose a methodology of collaborative lecture incorporating distance learning and will also provide the experimental evaluation by a prototype made for this experiment.

.

Keywords: collaborative lecture, distance learning, chat system, R/SW, Wiki

1 INTRODUCTION

With the growth of internet use in recent years, researches on the possibility of new educational methods including the internet technologies are actively practiced. One topic gathering attention on the standpoint of educational engineering is "Collaborative lecture". Varieties of collaborative lecture methodologies are being studied utilizing such supporting functions as bulletin boards, teleconferencing and chat rooms, etc.[1][2][3] These studies mostly focus on the grasp and control of the progress of conversations and discussions with the use of worldwide web functions.[4][5] The other one is "Distance Learning". Distance learning is the method which connects teachers and students, who are mutually separated at distant locations, with communication facilities. The distance learning is fundamentally divided into two types: one is the asynchronous type in which students access teaching materials via the internet anytime he wants to study, and the other is the synchronous type in which students interact directly with teachers real-time via communication facilities such as teleconferencing systems. Especially the latter is regarded as effective to activate the learning and to promote the creativity of students through the interaction with those of the other side. However, collaborative lecture systems tend to be used independently from distance learning, and very little work is being done on collaborative lecture methods that include distance learning.

The authors have proposed a distant-lecture system in which teaching materials are sent in advance to computers on the student side and teachers send editing commands to display the teaching materials to students.[6][7][8] We believed that if remote lectures could be included in collaborative lecture by students, the effectiveness of this collaborative lecture could be improved. This paper proposes a methodology for incorporating distant lectures in collaborative lecture and provides an experimental evaluation.

To evaluate this method we created a collaborative lecture prototype and performed evaluations within our department. The experiment was performed twice: the first time was to confirm and evaluate the basic functions of the system. Various problems centering around the system's chat functions were identified, and the result of the evaluation was unsatisfactory. The second experimental evaluation was performed after improvements to the system's chat functions, and repeated the evaluations addressing each of the problems.

2 A COLLABORATIVE LECTURE METHOD INCLUDING DISTANT LECTURE

2.1 Basic Approach

The general flow of collaborative lecture including distant lectures is described below. The whole methodology consists largely of three stages, the initial asynchronous learning stage, the distant lecture stage that comes next, and the final collaborative lecture stage.

(1)At the asynchronous lecture stage, the teacher sends the teaching materials with contents prepared for the course to the students in advance. The contents of the teaching materials are used when the distant lecture is

given, but it is also possible for the students to use them in their preparation.

(2)At the next distant lecture stage, the teacher sends commands to display the teaching materials, and explains the theme of the studies. Those receiving the lecture see displays corresponding to these commands. Once the explanation of a particular topic has been completed, the teacher sends a question command and checks the degree of understanding. If teachers perceive that understanding is inadequate, they will give additional explanation(s) and check again. When adequate understanding has been achieved, the lecture can proceed. The collaborative lecture stage begins once the lecture has been completed.

(3)At the collaborative lecture stage, students are divided into groups in which learning proceeds by way of conversations and discussions on the theme of the study and the joint creation of a document. The teacher watches the log of this process for each group, monitors the progress of document creation by group members and gives appropriate advice. If he perceives further instruction being necessary, they revert to the distant lecture stage and appropriate additional explanations are given.

Discussions on the study theme and document creation continue until the group is led to the right conclusions. Once the group has completed the study, the teacher confirms its completion, and the group uses the document(s) to prepare and submit a report. The study as a whole ends with the presentation of these reports. The teacher refers to the results of the study, including chat logs and the data on the information shared, in preparing feedback for the next lecture.

2. 2 Configuration of Collaborative lecture System

The system configuration is shown in Fig. 1. The collaborative lecture system consists of the software for editing teaching materials in real time (abbrev. to R/SW hereafter), the chat system, and the information sharing system.

R/SW has functions allowing commands to be sent during the lecture that cause the content of the teaching materials to be displayed to students. In addition to these content display commands, there are also commands allowing the teacher to check the degree of student understanding. There are also functions to analyze the results of queries.

The chat system uses NetMeeting chat, with tags that provide graphic identification of the kind of content that is being sent. This eliminates the misunderstandings that can easily arise in text-only chat sessions, and assists in judging what kind of discussions were being carried out by the group members as they followed the chat log. These tags also facilitate the investigation and analysis of the learning process from the chat log. The tags used in the proposed method are as follows.

(1) Proposal : Identifies proposals
(2) Question : Identifies questions
(3) Explanation: Response and explanations in reply to questions
(4) Agreement : Expresses agreement
(5) Opposition : Expresses opposition
(6) Other : Content not among the above

Because we considered that discussions would proceed by a repeated process of reaching agreements after exchanges of opinion between the participants, we set up tags to express "agreement" and "opposition" with "question" and "explanation" for questions and answers. Since conversations would not be confined to discussions, we also prepared an "other" tag.

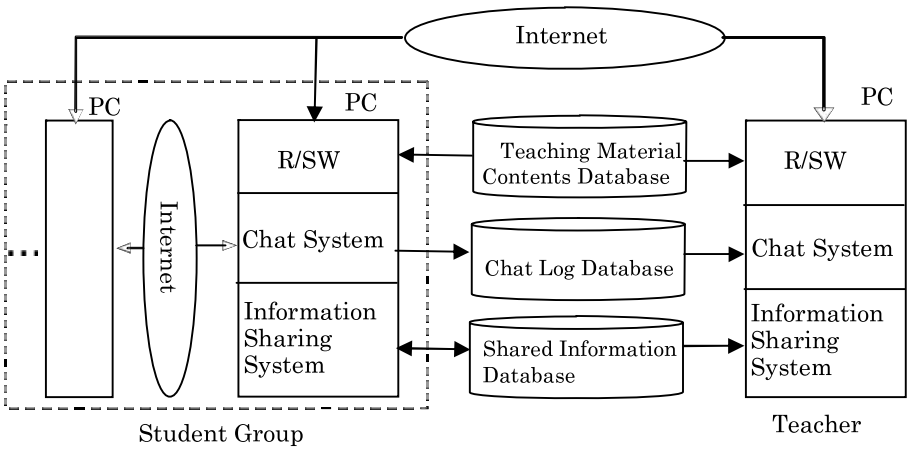


Fig. 1 System Block Diagram

We used a Wiki[9] as the system for sharing information. A Wiki is a web contents management system that facilitates access via a web browser, and allows freedom in issuing and editing pages.

2. 3 System Support

In order for students to follow the learning theme and to ahieve the objectives of the study, a document giving the results of their studies is prepared using discussions that proceed via chat and an information-sharing server. In this process, logs are accumulated of the chat conversations and the changes made to the document on the information-sharing server. Real time support during the lecture time is provided by analysis of the data on the server. R/SW is used to grasp the students' answers to questions, and provides the additional factor of distant lectures that are responsive to the students. It is very important, for the proper support of this kind of learning, that the process of distant collaborative lecture should be analyzed, and this is done using the log of conversations and document creation.

2.4 Grasping the Students' Responses

In order for the degree of understanding to be properly assessed in synchronous learning, R/SW commands are used to question understanding and R/SW analytical functions are used on the resulting answers. Here, the degree of student understanding is assessed on the basis of multiple-choice Q&A forms and single-question answer forms for 20- to 40-word answers. Selection is performed in response to the teacher's questions by means of branching "radio" buttons.

The responses accumulated on the teaching side are checked to grasp the degree of understanding. The answers are totaled, and the average numbers of those selecting a particular branch option are ascertained. This provides a quantitative assessment of the overall level of understanding. The single-question answers are scanned for the keywords that they should contain to assess their correctness. For example, if asked what an interpreter language is, the answer is assumed to be correct if it contains the words "interpret" or "execute.".

3 EXPERIMENTS AND EVALUATION

3. 1 The First Test, Its Evaluation

3.1.1 The Experimental Environment within the Department

This experimental test was performed with ten fourth-year students in the department. They were divided into two teams, A and B, for the studies. Because the experiment envisages students at distant locations, the students were required not to sit next to one another in the lecture room, and not to communicate face-to-face. All students possessed notebook computers, and were connected via a wireless LAN. The students' computers had pre-installed R/SW, NetMeeting (used for chat) and the Wiki used as the shared information server.

3.1.2 Experimental Schedule and Study Theme

The experiment was performed five times in total, of which first to four sessions took 90 minutes each time. In a fifth and final session, each team announced its results. The theme of the collaborative lecture for this experiment was "An Investigation of the Suitability of Distance Learning Using the Internet for the Regions Surrounding the Takla Makan Desert and, if Suitable, of Problems Arising," and after the teacher lectured on the theme of the study, the group commented collaborative lecture as discussions continued.

3.1.3 Results and Evaluation

For this experiment, we performed analyses of the tags, the chat conversation flow, the Wiki information flow, the distant lecture itself, the questions and answers, the students' announced results, and the results of a questionnaire and opinion survey. As a result, we were able to confirm the overall flow of the collaborative lecture process including distant lectures, but the following problems were identified. (1) Too much use was made of the "Other" tag (35.7%), making analysis of the discussion difficult; (2) the widely differing time lags between discussions and document preparation tended to disrupt discussion; (3) it was difficult to know the situation of other students, etc. We accordingly decided that satisfactory results could not be expected if the experiment were to be continued in its current form, and made improvements before carrying out the next experimental test and evaluation.

3.2 System Improvements

3.2.1 Improved Chat System

For this research, we developed a chat system with tags that would give simultaneous information on the state of discussions and document preparation. We have called this "semantic chat." A typical semantic chat screen is shown in Fig. 2.

The chat log with tags is shown in window (1), the tag selection list in window (2), the log of document creation in window (3), the area for entering contributions in window (4), and the list of participants in window (5). Meaningful tags are selected from the list in (2), and they are also allocated to function keys for selection and input without having to remove hands from the keyboard.

The chat log is on the left, but in the window on the right, Wiki changes are logged along the same time line.

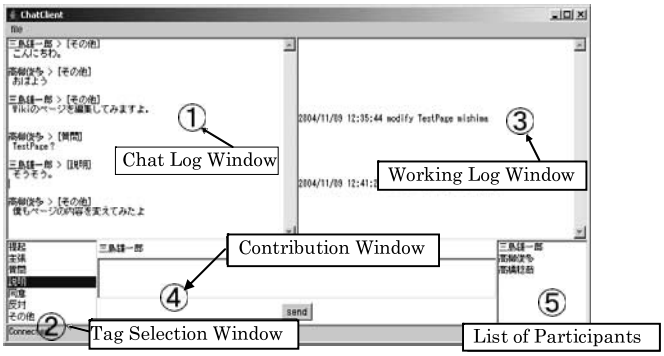


Fig. 2 Typical Semantic Chat Screen

This means that as collaborative distant lecture proceeds in parallel with document preparation, the progress of both discussions and preparation can be grasped simultaneously. We believe that this kind of environment, by providing a log of progress in documentation, will enable the teacher to distinguish between those students who are and are not participating in the study, whether or not they are actually participating in the active give-and-take of discussion.

3.2.2 Short Supplementary Lectures During Collaborative lecture

In this experiment, we tried the method of having the students attempt to resolve among themselves any questions arising from the study theme during the collaborative lecture process. However, analysis of the content of the discussions and the jointly prepared reports revealed that to some extent supplementary lectures from the teacher were necessary. In the next experiment, we therefore arranged to ascertain the student response and to provide appropriate supplementary instruction.

3.3 Experiment using the semantic chat system, Its Evaluation and Consideration

3.3.1 Test Schedule and Content

In the next experimental test, nine of the department's fourth-year students participated on two occasions. Nine of the participants in the previous test were divided into two teams, Team A with four members and Team B with five. The study theme this time was different; "A Review of Policies to Encourage More Japanese Tourists to Visit the Autonomous Uighur Region of Xinjiang Province on China's Silk Road." There were two group discussions. One concentrated on what could be done to bring more Japanese tourists to China's Silk Road and the Autonomous Uighur Region, and the other concentrated on proposing appropriate measures to solve the problems involved. Participants were also pre-tested by e-mail before the experiment.

On the first occasion, the teacher gave the lecture on the study theme, then checked the degree of understanding, gave a supplementary lecture, and then rechecked understanding.

On the second occasion, collaborative lecture took place, with checks of student interest, additional lectures, checks of comprehension, etc., and finally each team brought its document to a conclusion and submitted a report. The experiment was followed by a test and a questionnaire.

3.3.2 Experimental Results and Evaluation

On the first occasion, the students' reactions were ascertained after the distant lecture. Check points were displayed on the R/SW communications screen, and student understanding during the distant lecture was assessed by asking questions with the answers to be selected from four alternatives. This check was repeated four times. The results achieved by the students were assessed and additional lectures were given where this appeared to be necessary, and the same four questions were asked to check understanding. According to the checks of understanding before the additional lecture, the number of students who answered correctly was 7, 7, 8, 7, while the number of those answered incorrectly was 5, 4, 2, 4, and the number of those not answering was 0, 1, 2, 0. This resulted in the percentage of right answers being 60.42%.

Compared with this, the number obtained after the additional lecture increased to 12, 11, 11, 12 for those answered correctly, while the number decreased markedly to 0, 0, 0, 0 for those answered incorrectly and to 0, 1, 1, 0 for not answering. This improved the percentage of right answers to 95.83%. From this result, it is clear that student understanding improved considerably.

On the second occasion, during the collaborative lecture, the teacher analyzed the semantic chat log and the document, and realized where students were experiencing problems and where more detailed explanations were necessary. First, the degree of student interest in this possibility was checked. The check was performed twice for Team A and Team B. On the first check, the number of students who were interested and needed supplementary explanations was 4 for Team A and 4 for Team B equally. Then the number who responded no need for supplementary explanations was zero for both teams, and the number who answered either would do was 1 for Team A and zero for Team B. On the second check, the numbers of students who said necessary, unnecessary, either for supplementary explanations were 4, 1, 0 for Team A, and 3, 0, 1 for Team B. From this result it is clear that there was a need for supplementary lectures during collaborative lecture.

Next, the necessary supplementary lecture was given and student understanding in additional instruction was

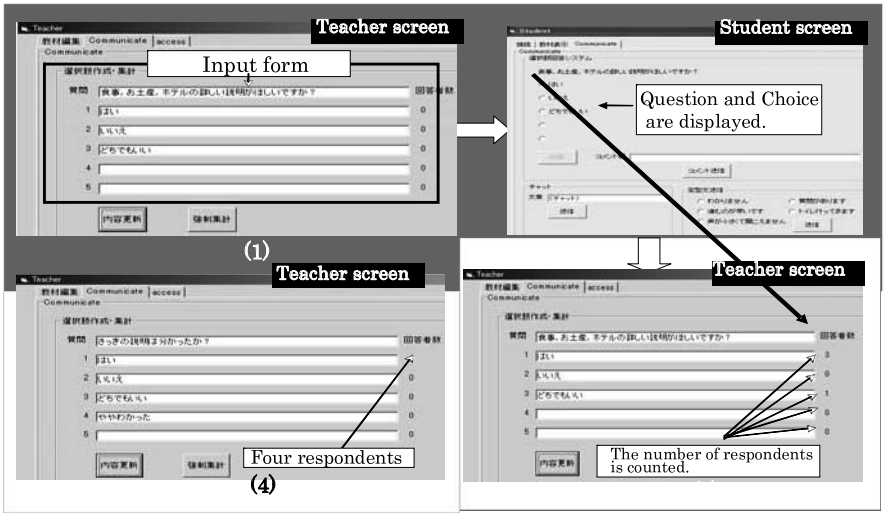


Fig. 3 Screens for Checking Student Understanding and Interest

checked again, with the results shown in Fig. 3. Parts (1), (2) and (3) of Fig. 3 are the screens on the teacher's side for checking the degree of student interest before giving a supplementary lecture and Part (4) is the screen on the teacher's side for showing the statistics for the degree of student understanding after the supplementary lecture. Part (1) shows the teacher-side selection of questions, (2) shows the student-side questions and their selection of answers, (3) shows the teacher-side statistics for the degree of student interest, and (4) is the teacher-side display of student understanding. The example was "Would you like explanations of the meals, souvenirs and hotels available in the Uighur Autonomous Region of Xingjiang Province?" The students answered by selecting from three options: "yes," "no" and "either" (i.e., "don't care"). After the supplementary lecture was given, their understanding was checked by the question "Did you understand this explanation?" and having them select one of four responses. Clearly, almost all of the students claimed to understand it.

3.3.3 Considerations

(1) Log Analysis and Semantic Chat

In the course of experiments of collaborative lecture including the distant lectures, we compared the frequency of the "others" tag usage in the first collaborative lecture experiment when the NetMeeting chat was used with that in the second experiment when the semantic chat system was used. In the NetMeeting chat, frequency of the "others" and non- "others" tag usage was 35.7% and 64.3% respectively. In contrast, the frequency of usage changed to 23.9% and 76.1% in the Semantic chat. This means that when the semantic chat system is used, the "others" tag is used some 11.8 percentage points less than when NetMeeting was used. This suggests that the method of associating tags in the semantic chat system has become considerably more convenient, and that the students had become more familiar with their usage.

A questionnaire survey was carried out after the study. The contents were divided into 15 categories, each with five levels of response. It also allowed for open-ended expressions of opinion. In response to the questions "Was the semantic chat system effective?" "Did the semantic chat window showing the updating of Wiki information prove useful?" and "Was the supplementary lecture helpful?", many students mentioned as positive factors that the semantic chat system was good in that it had simplified the input of tags, and that the Wiki update window facilitated progress by eliminating wasteful confirmations.

However, there were also students who indicated that although the display of information on the work of other students during the discussions was effective, there was also a need to display the status of the discussions with other students while the work was proceeding. We call this kind of situations in the collaborative lecture as "awareness". We recognize this kind of mutual attention is important to enhance the sense of cooperation among the students and how to casually convey this kind of information remains to be solved in the future.

(2) Analysis of Supplementary Lectures

From the discussions in the Experimental Results and Evaluation, it is clear that the supplementary lecture improved student understanding, and that supplementary lectures are also necessary during collaborative lecture. The results of tests of understanding performed after the supplementary lectures show that almost all students had understood their content.

(3) Pre-Test and Post-Test Analysis

The same test of the study theme was performed both

before and after the collaborative lecture. In the pre-test, average scores of Team A and Team B were 22.75 and 17.33 respectively, while in the post-test, they increased to 69.5 and 76.66. Resulting differences were +46.74 for Team A and +59.33 for Team B. The average scores show considerable improvements before and after the collaborative lecture. In terms of actual test scores, all students achieved higher scores.

(4) A Comparison of NetMeeting and Semantic Chat

NetMeeting was used for discussions in the first experiment and semantic chat was used in the second. Semantic chat dispensed with the need to enter tags manually, and we expected this would affect the number of contributions. The number of contributions per unit time in the first and second experiments is shown in Fig. 4.

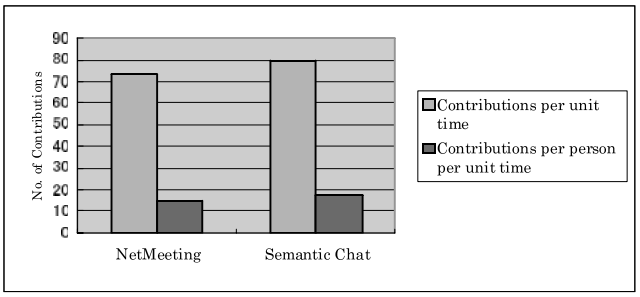


Fig. 4 Comparison of the Number of Contributions made using NetMeeting and Semantic Chat

The comparison of NetMeeting and semantic chat given in Fig. 4 shows that both the total number of contributions to the discussion made per unit time, and the average number of contributions made by individuals per unit time, are both higher for semantic chat. This appears to be because dispensing with the need to enter tags by hand has encouraged more contributions.

4 CONCLUSIONS

This paper proposes a method of collaborative lecture that incorporates distant lectures and provides its experimental evaluation. The first experimental evaluation was performed for collaborative lecture with the theme "An Investigation of the Suitability of Distance Learning Using the Internet for the Regions Surrounding the Takla Makan Desert and, if Suitable, of Problems Arising." As a result of analysis of the study process we encountered several problems such as inefficiency of tag usage, mismatch of synchronization of discussions and document preparation, difficulty of knowing other students' situation etc. Having judged that it was impossible to continue the experiment without any change, we developed the semantic chat system which was able to grasp discussions and documentation at the same time.

In the next experimental evaluation, the study theme was "A Review of Policies to Encourage More Japanese Tourists to Visit the Autonomous Uighur Region of Xinjiang Province on China's Silk Road." As a result, improvement of the tag usage was observed and which had influence to the increase of contributions. However a new problem of "awareness" was identified in this method.

Future issues to be resolved are the establishment of the parameters by which the effectiveness of studies performed via collaborative distant lecture can be assessed. We also plan further practical testing and improvement of the developed system, and intend to perform experimental evaluation of the time lags between discussion and document preparation.

REFERENCES

[1]The Advanced Learning Infrastructure Consortium (ALIC) (Ed.) "*On Collaborative Learning*", the annual White Paper on e-Learning 2003/2004 edition, P295, Ohmsha Ltd., Tokyo, 2003. (in Japanese)

[2]"The Virtual UNIVERSITY",Edited by Kevin Robins and Frank Webster, OXFORD UNIVERSITY PRESS, 2002.

[3]"Virtual Environments for Teaching & Learning",Edited by L.C.Jain etc,World Scientific,2002

[4]Akiko INABAetc, "*An Intelligent Supporting of Discussion for the Distributed Cooperative Learning Environment*" Transactions of the Institute of Electronics, Information and Communication Engineers (IEICE), A Vol.j790A No.2, pp207-30, 1996. (in Japanese)

[5]Keisuke YAGI etc, "*A Novel Distance Learning System for the TIDE Project*" Transactions of the IEICE, D-II Vol.j84 – DII No.6 p.1139, 2001-6. (in Japanese)

[6]Yuichi MISHIMA,Tomoo INOUE,etc, "*Activity-aware semantic chat system based on a study of learning process in distance collaborative learning*" Information Processing Society of Japan Special Interest Group (IPSJ SIG) on Groupware and Network Services Workshop 2004, pp81-86,2004. (in Japanese)

[7]Dilmurat Tilwaldi etc "*A Real-time Editing Method of Teaching Materials in the Unified Synchronous/Asynchronous Distance Learning*" IPSJ SIG Technical Report DPS-113, PP385-386,2003-6. (in Japanese)

[8]Toshiya TAKAHASHI etc, "*A Proposal of An Education Support System with Functions of Real-time Editing Contents based on WebCT*" IPSJ Workshop on Multimedia Communications and Distributed Processing, pp143-148,2004-12. (in Japanese)

[9]Hisayoshi ITO, "*An Attempt of Information-sharing within a Laboratory on Wiki Clone*" IEICE Technical Report, Vol.103 No.226, pp13-18, 2003-7. (in Japanese)

Proposal and Evaluation of Pictograph Chat for Communication

Jun Munemori, Sumika Ohno, Takashi Yoshino, Junko Itou

Faculty of Systems Engineering , Wakayama University, Japan
{munemori, yoshino, Itou}@ sys.wakayama-u.ac.jp

Abstract **-** In Japan and China, pictographs have widely spread to add nuance to mails of mobile phones. We have developed a pictograph chat system, which can communicate each other using nothing but pictographs. We prepared 550 pictograph symbols. We applied the system for communication to 3 groups, which consist of the intimate friends group, the strangers group, and the foreign students and Japanese students group. We have carried out experiences 18 times. We report the results of the experiments as below. (1) The subjects understood over 70% of the content of the chat. (2) There were no difference between foreign students and Japanese students about the context of the chat. Japanese students tended to make the same context as foreign students.

Keywords: CSCW, Groupware, Pictograph, Chat, Cross-cultural communication

1 INTRODUCTION

The communication of information by E-mail, chatting, and electronic bulletin board has become widespread by the spread of networks. Moreover, communication can be easily done using MSN Messenger [1] etc. with a text base. In addition, face marks and pictographs have appeared, and are used to convey feelings [2].

Language becomes a barrier if we think about communication between different countries, and if a common language is not understood, it is difficult to communicate through conversation on a text base. Moreover, to learn a foreign language requires considerable time. Therefore, attention was given to pictographs that are used to convey feelings and slight nuances based on the idea that communication is possible if pictographs are used, even if a language such as English is not understood[1].

To add pictographs to chatting when experimenting on a teleconference between Japan and China [3], a recognition investigation of pictographs was conducted with postgraduate students from Japan and China (eight from each country). 112 pictographs were made for, and shown to, the subjects, who then evaluated their suitability for the meaning. As a result, between the Chinese and Japanese subjects, only 4 pictographs differed greatly in recognition (a school, house, motor sports, and rice ball). The pictograph for “school”, for example, looked like a regular building. Though understandable to Japanese subjects, the Chinese did not see it as a school because Chinese had a grand image like a castle for school. "Rice ball" is not seen in China.

As a result, it was seen that the recognition of the pictograms was almost unchanged between the Japanese and Chinese subjects. We then sought to determine whether subjects could understand even if sentences were made and chatted only by pictographs, and developed a system in which only pictographs were used for chatting. This system is a chat system that can send and receive messages made only from pictographs. 550 pictographs were prepared including ones to do animation. This system was actually applied to chatting among the groups of "Friends of a friend", "Persons who did not usually talk ", and "International students and Japanese students". From the results of the experiment we considered whether communication that used only pictographs was possible.

Chapter 2 explains the pictograph chat system. In Chapter 3 we explain the understanding level that is the standard of the evaluation of the experiment that uses this system. Chapter 4 shows the experiment that uses this system. Chapter 5 describes the experiment results. Chapter 6 is a discussion. Chapter 7 describes an additional experiment. Chapter 8 is the summary.

2 PICTOGRAPH CHAT SYSTEM

2.1 Composition of system

This system was developed in C# language using Microsoft Visual Studio.NET 2003. It is a program of about 1200 lines. The system consists of a chat log screen with a pictograph input screen where pictographs are selected and sentences are made.

2.2 Functions of system

Figure 1 shows the pictograph input screen. The chat log screen is shown in Figure 2. A pictograph can be added from the pictograph selection screen to the input field with one click. A new pictograph can be input by selecting the pictograph in the input field, and inputting a new pictograph from the pictograph selection screen between the pictograph that exists originally in the input field and the pictograph selected by clicking.

[1] The work reported in the paper was partially supported by Japan
Society for the Promotion of Science (JSPS), Grant-in-Aid for Scientific Research (B) 18300043, 2006 and 20300047, 2008.

Figure 1: Pictogram input screen.

Figure 2: Chat log screen.

Moreover, the selected pictograph can be deleted by clicking the Select Image Clear button, and pushing the Back Space key and the Delete key. When the auto animated cartoon button is checked, all pictographs of possible animation begin animation.

Pictographs were the original 255 images, 295 made of PIC-DIC [4], and 550 images in total were prepared. An original pictograph is one in which a newer pictograph was added referring to the pictograph used for cellular phones. Permission was obtained from Godai Embody Co., Ltd. for the use of PIC-DIC pictographs. Original pictographs are displayed in color, while PIC-DIC pictographs are displayed in black and white.

3 UNDERSTANDING LEVEL

The meaning of the chat that a subject wrote by himself/herself and the other party's interpretation of the chat were compared, and the degree of understanding of each subject was calculated. The understanding level includes the understanding level in one direction (one person's understanding level) and interactive (understanding level of the conversation of two people).

The calculation of the understanding level was shown by pulling the amount not understood from the whole. The following three patterns are used to determine the part not understood. The setting assumes that a conversation of 20 lines is comprehensible. Ten of the lines are the remarks of subject A, and the other ten lines are the remarks of subject B. We assume to be comparing here the interpretation of Mr. B of the remarks of Mr. A.

(1) The interpretation is quite different.

The sentences of one line assume that the understanding level is 0. The understanding of one line's worth is pulled from the whole. For instance, when Mr. B is doing a quite different interpretation by three lines at this time, the understanding level of Mr. B becomes an understanding level of 70%(7/10) that pulls 30%(3/10) from 100%.

(2) The interpretation of several pictographs of one line is different.

Figure 3: An example of the pictograph chat (The interpretation of several pictographs of one line is different).

For instance, in the example of Figure 3, Mr. A is making the remarks, "I made and ate pork with rice at home". Mr. B is interpreting, "I made and ate a meal at home ". In a word, the amount of the pictograph of enclosed "Pig" is not interpreted with the full six pictographs. In this case, 1/6 is pulled from the understanding levels of these sentences. For instance, the calculation type when Mr. B doesn't understand two lines at all, and one line is interpreted like this is as follows.

$$(7/10) + (5/6) \times (1/10)$$

Therefore, the understanding level of Mr. B becomes about 78%.

(3) The interpretation of one pictograph looks similar.

Figure 4: An example of the pictograph chat (The interpretation of one pictograph slightly looks like).

For instance, thinking about the example of Figure 4. Mr. A is making the remarks, "Will it be cold tomorrow?" At that time, Mr. B assumes that he interpreted, "Will it snow tomorrow?" The interpretation of "Snowman" enclosed in a word full among six pictographs is different. However, "It is cold" and "It snows" look similar. Only 1/2 of these understanding levels of one pictograph is pulled from the whole and calculated. In a word, the calculation type when Mr. B

doesn't understand two lines at all, for instance, and one line is interpreted like this is as follows.

$$(7/10) + \{ (5/6) + (1/6) \times (1/2) \} \times (1/10)$$

Therefore, the understanding level of Mr. B becomes about 79%.

For instance, "Understanding level of the conversation of two people (understanding level of interactive)" when Mr. A doesn't understand the remarks of Mr. B at all in four lines, and Mr. B interprets as (2) is as follows.

$$(13/20) + (5/6) \times (1/20)$$

Therefore, the understanding level of the conversation of two people (interactive understanding level) becomes about 69%.

4 EXPERIMENT

Two subjects experimented using personal computers through LAN in separate rooms. It was not possible to communicate verbally at all. The subjects were all students at Wakayama University. One was a third-year student of the Department of Economics, four were third-year students of the Faculty of Systems Engineering, 18 were fourth-year students, three were first-year graduate students of the system engineering research course major, and two were second-year graduate students. Similarly, two were third-year students, 3 of the international students were first-year graduate students, and one was a second-year doctoral student. The breakdown of the international students is four Chinese students, one Malaysian, and one Vietnamese. The subject combinations were as follows.

Experiment 1: Japanese with good relations

Experiment 2: Persons with whom Japanese don't usually talk mutually

Experiment 3: Japanese and international students

Subjects were 36 in total, with six pairs for each of the three patterns. Experiments were done two times for each second-year graduate student (experiment 2 and experiment 3). Otherwise, it was only one time for each. Whenever a one-line remark went out to the chat, the subject wrote the meaning of the remark of the other party and his remark while experimenting.

5 EXPERIMENTAL RESULTS

5.1 Example of each experimental chat

Figure 5 shows the experiment example of the screen. The result of the experiment 1(Japanese with good relations) is shown in Figure 6, and the result of the experiment 2(one about which the Japanese doesn't usually talk mutually) is shown in Figure 7 and the result of the Japanese and the international student is shown in Figure 8. The friend with a good relation of the fourth grader is talking with Figure 6 concerning meal and the graduation thesis. The mark of the first animal of Figure 6-8 shows the speaker.

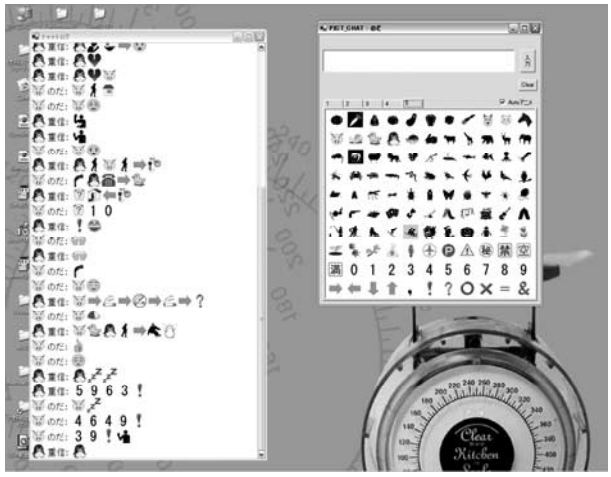


Figure 5: Example of screen of pictograph chat system.

"Did you eat lunch at school?"

"I ate at school. Did you come to the school after eating at home? "

" I made and ate pork with rice."

" Is your thesis advancing? I am stumbling. "

" I seem also to be in a slump. I cannot hit on what I need to write."

Figure 6: Chats of the experiment 1 (partly).

A conversation the same as usual is done.

Figure 7 is a conversation between Japanese who don't usually talk.

"Do you like to sing?"

"Yes. "

""I love it."

"Work part-time by karaoke. "

"What?"

" I am working part-time in a karaoke shop. "

「Oh!」

Figure 7: Chats of the experiment 2 (partly).

The conversation to start getting to know the other party is done. Because the other party did not understand, it would have been written, "I am working part-time in a karaoke shop." Once, the other party's understanding is obtained by explaining again in detail in the pictograph.

Figure 8 is a conversation between a Japanese and an international student.

"Do you like karaoke?" (Japanese)

" I do " (international student)

"I like it" (Japanese)

"What kind of singer do you like?" (international student)

Correct content: "I like Mr. Children. "(Japanese)

Wrong content: "I like Boys and men. " (international student's interpretation)

"I like it. " (international student)

"Really"(Japanese)

Figure 8: Chats of the experiment 2 (partly).

Another singer (Mr. Children and Boys and men) is mutually imagined though the conversation was going to have been approved by both people.

The following is understood from the results of the questionnaire taken after the experiments of application and the chat log.

5.2 Number of distribution of pictographs by subject

Figure 9 shows the distribution of pictographs for each user. The average number of pictographs in experiment 1, experiment 2, and experiment 3 is 44. About 3.2 pictographs on an average line (one remark) are used. The maximum number of pictographs on a line (one remark) is 10 in experiment 1, 10 in experiment 2, and 12 (though, if it is excluded, eight are the maximum because the same pictograph is repeated) in experiment 3.

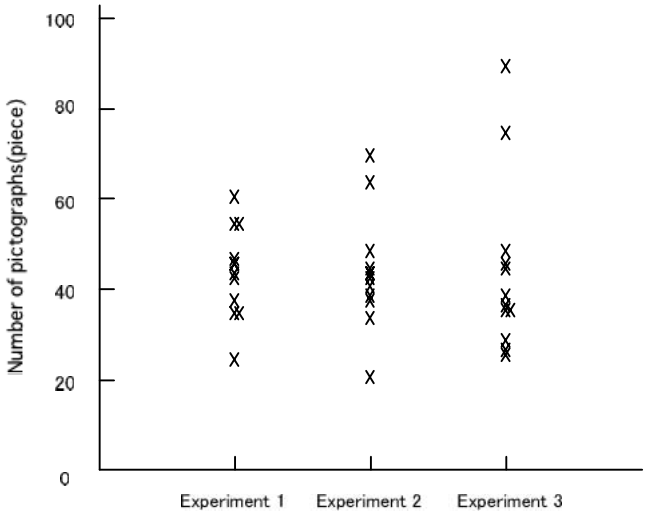


Figure 9: Distribution of pictographs for each user.

5.3 Understanding level of conversation

The mutual understanding level of the conversation of each experiment was calculated by comparing the meaning of the remark that the subject had written. The average of the understanding level of experiment 1 is 73%. The average of the understanding level of experiment 2 is 81%. The average of the understanding level of experiment 3 is 78%.

5.4 Questionnaire results

Part of the results of the questionnaire of five-point scale evaluation (Table 1) and opinions are shown. "5" is the highest score and "1" is the lowest.

Table 1: Results of the questionnaire.

Questionnaire items	Evaluation
When the pictograph is clicked once with this system, the pictograph is added to the input field. Was the operation easy?	4.6
Was sentence making easy?	2.0
Was the auto animated cartoon function convenient?	4.2
Could the meaning of an individual pictograph be understood?	3.6
Was a target pictograph easily searchable?	2.4
Was there a target pictograph?	3.1
Did you understand what the other party said?	3.8
Do you think that you were able to underatand the other party?	3.7
Doyou think that you can conduct a conversation chatting only with pictographs?	3.3
Was this experiment interesting?	4.6

Moreover, examples of the description-type results of the questionnaire are shown below.

- In what situation do you think that this pictograph chat can be used?

It is possible to use it cross-culturally.

It is possible for a child to use it.

It is possible to be used among companions usually.

It is possible to use it for communication with a handicapped person.

- Please write if you have an opinion about this system.

You should introduce the grammar.

As for pictographs, easy communication can be done even cross-culturally. It is far better than characters when it is used well.

If pictographs are expanded to do a complex conversation, it takes time to look for them. However, if one manages to do it, it is possible to use it well.

[Opinion]

・I think that I was able to understand more than I had expected.

・It is interesting to decipher what the other party wants to say.

・The animation is lively and happy.

・It takes time to look for a target pictograph.

・Aren't some rules needed?

・The pictographs are few (Proper nouns cannot be expressed).

6 DISCUSSION

6.1 Difference by experiments in understanding level

The understanding level of experiment 1 is 73%, which is lower than the understanding level of experiment 2 and experiment 3. Two subjects in experiment 1 often talked usually, and were well-acquainted with each other. Therefore, they talked as an extension of their usual conversation, and the content of the conversation became deep. Much of the content was a continuation of the conversation of two people on the day before the experiment on topics like research, graduation travel, Valentine's Day, etc. However, there were no pictographs of proper nouns for the content of deep conversation. Therefore, we thought that it caused a gap in their interpretations, and it led to such a result. There was no topic for "Ice-breaking" that was used to be mutually friendly.

The understanding level of experiment 2 (81%) is a little high compared with the understanding level of experiment 1 and experiment 3. Two subjects of experiment 2 were two people who might not speak to each other especially. Ice-breaking topics, for example the weather, appeared in all six experiments. The sentence composition is simple. The subjects understood most though there was a slight mix up in individual pictographs. Therefore, it is thought that it led to such a result.

The understanding level of experiment 3 is 78%, which is higher than the understanding level of experiment 1, but lower than the understanding level of experiment 2. The subjects of experiment 3 are an international student and a Japanese student. There were a couple of subjects who did not usually speak to each other at all, and there were five pairs who met for the first. Therefore, there appeared in all of the six times in the ice-breaking conversation "What time did you sleep yesterday?" "Do you like skiing?" "Do you like soccer?" "Do you like karaoke?" "Do you like certain singer?" as well as in experiment 2. Moreover, the other party's country of origin was occasionally heard. However, the name of a country's not being given happened because in that case, there was no pictograph for the proper noun, and the topic changed suddenly. However, a panda pictograph was used to depict China. There was a conversation like "What do you like?" In this case, they were not able to talk well due to absence of expressible proper nouns. The conversation level reached an understanding level that was higher than experiment 1, because it was not different from experiment 2. We think

that experiment 3 had a result that the understanding level is lower than experiment 2 due to the problem that there were no proper nouns though we tried to get to know the other party in detail. Moreover, international students alone did not differ in their evaluation scores on the questionnaire (Table 1). We think that they used the system in the same sense as the Japanese.

6.2 Difference of description of Japanese and international students

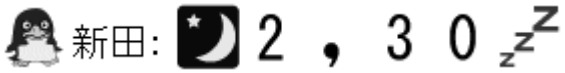
We consider whether there is a difference in the sentence composition between Japanese and international students from the chat log regarding the composition of sentences. Figure 10 shows a comparison of the conversations "What time did you get up today?" and "What time did you sleep yesterday?" that had come out while subjects did a lot of talking. Seeing this figure, it was understood that the Japanese and international students also include persons who combine pictographs of action after the pictograph of time and those who combine pictographs of time after pictographs of action.

What time did you get up today? (Japanese)

What time did you get up today? (international student)

I was sleeping until 6:00 yesterday. (Japanese)

I slept at 4:00 yesterday. (international student)



I slept at 2:30 at night. (Japanese)

I slept at four o'clock of this morning (international student)

Figure 10: A comparison of sentence construction.

6.3 Extraction and correspondence of problem

There were various devices when it was not possible to correspond directly to the pictograph. For instance, China is shown by using a panda pictograph because there is no pictograph named China in experiment 3 for the question "What country do you come from?" Moreover, there was a device to make it read by an adult pictograph and baby pictograph as Mr. Children (Figure 8).

From the results of the questionnaire, 69% said the pictographs could be understood, and 56% said there were appropriate pictographs. Especially, the lack of proper nouns was pointed out. In addition, the demand for needed pictographs were the following.

・"Yes" and "No", seasonal pictographs, and pictographs that show other people.

・Time expression such as "Former", "Old times", "Soon", "This morning", "Month", "Day", and "Year"

・ What, when, where, who, how, and adverbs, conjunctions, proper nouns, units, and signs.

・Proper nouns

We have to deal with the lack of proper nouns. Hieroglyphic Tompa characters [5] have a lot of proper nouns. It is possible to show them basically by combining meanings. Japanese is expressible by Tompa characters. For instance, you may express the proper noun "Ishikawa (Stone and river)" by combining characters that show "Stone" and "River". It is uncertain whether this meaning can be communicated in a conversation between a Japanese and a foreigner. As for hieroglyphs [6], one alphabet character corresponds to one pictograph as a proper noun. Subjects were Japanese, Chinese, Malaysian people, and Vietnamese people. Because these subjects can at least read the alphabet, it seems that it is necessary to prepare proper nouns to write by the alphabet.

6.4 Related works

It is a project of the NHK South Pole kids project [7], and there is a pictograph chat system for children all over the world to do communication using only pictographs. Concerning this system, pictographs which may be lined up to eight are expressible by the chat system of the Web base. It doesn't have an animation function. Similarly, research involving communication with children in different countries using pictographs was done [8]. However, it is a system using not a real-time chat but a mail base.

There is sign language as a method with the possibility that communication can be done excluding conversation. However, there are a lot of dialects of sign language by country. The sign language that shows Japanese is different from that which shows Chinese. At present, there is international sign language [9] common to all parts of the world though it is not so general. The comparison with this is a problem for the future. Moreover, a person in the sphere of Chinese characters can communicate in writing. The comparison with the Chinese character is a problem for the future.

7 ADDITIONAL EXPERIMENTS

That the sentence structure (order of the pictograph) is different depending on the person has been understood from the experiment. However, there is a possibility that

the sentence structure is controlled by written sentences. Then, the same sentences were shown to the Japanese and Chinese who lived in China in their respective mother tongues as an additional experiment, and it was written using pictographs. The Japanese subjects were 20 students at Wakayama University, and the Chinese subjects were 11 staff members in the Institute for Digitization of the Palace Museum Heritage in Beijing. We didn't use the system, but administered a paper questionnaire form. We examined how many pictographs or face marks to use beforehand. We examined whether there was a pictograph or a face mark used in ten recent mails. As a result, subjects who used 0 to 2 pictographs or face marks were 9 people, 3 to 6 were 4 people, and seven or more were 7 Japanese subjects. Subjects who used 0 to 2 pictographs or face marks were 7 people, 3 to 6 were 2 people, and seven or more were 2 Chinese subjects.

< I like bananas>

Japanese

S + V + O -> 70%

Chinese

S + V + O -> 37%

Japanese

S+O+V-> 20%

Chinese

S+O+V-> 27%

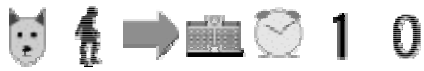
Figure 11: Results of experiment (S+O+V).

Figure 11 shows an example when the Japanese and Chinese write sentences "I like bananas" using pictographs. When these sentences were written with pictographs, the ratio that the Japanese wrote in order of subject + verb + object was 70%, and the ratio for Chinese was 37% though these sentences became the order of subject + object + verb if it was written in Japanese. The Japanese was 20% in the order of subject + object + verb that was the original Japanese order of writing, and 27% in Chinese. The total does not reach 100% because there are examples that cannot be classified like this as the pictograph is described by only one character. Next, the example when the Japanese and Chinese write sentences "I go to school at ten o'clock" with pictographs is shown in Figure 12. If these sentences are written in Japanese, it becomes the order of subject + complement + complement + verb. The order of the writing of (subject) + verb + complement + complement is 90% in the Japanese when these sentences are written with pictographs, and 64% in Chinese. The ratio that the Japanese wrote in order of complement + verb + complement was 10%, and 0% for Chinese. The total does not reach 100% because there are examples that cannot be classified like this as the pictograph is described by only one character.

< I go to school at ten o'clock>

Japanese

90%



Chinese

64%

Japanese

10%

Chinese

0%

Figure 12: Results of experiment (S+C+C+V).

When hieroglyphic [6] Tompa characters [5] are seen in the word order, subject + verb + object it is basic, though it is not a pictograph (There are exceptions, as well). Moreover, the word order is verb + subject + object in the hieroglyph. The verb comes ahead of both objects.

Pictographs are often written in the same order in Japanese and Chinese with a different original word order. Moreover, the understanding of each pictograph hardly changes. It seems that the understanding level that exceeds 70% is obtained if it limits it to daily conversation. The questionnaire results concerning how many pictographs to use with a cellular phone and the experiment results are compared. There is nothing said that he can't write with pictographs because the age is high. Moreover, it was understood that a person who uses pictographs well is not necessarily good at communication only with pictographs.

8 CONCLUSION

We have developed a system for chatting with only pictographs using 550 pictographs. This system was actually applied to "Japanese with good relations", "Persons who did not usually talk mutually", and "International students and Japanese students" and it was applied to 36 a total of people. The following was understood as a result.

(1) It is possible to communicate 70% or more even by sentences only of a combination of pictographs if it is a simple conversation.

(2) There is no difference of the evaluation value in the difference of the sentence composition between Japanese and international students when talking by combining pictographs, and in the questionnaire. Japanese tend to write pictographs in the same order as the international students.

(3) The correspondence such as preparing more pictographs to express proper nouns when friends have a deep conversation, and the ability to input proper nouns, is necessary.

We think it is possible to communicate enjoyably using1 pictograph chatting. However, we admit the need for improvement of the lack of pictographs showing proper nouns. And, we think it is an effective system as a new method of communication. Moreover, we think that between friends and persons meeting for the first time it is effective to communicate only by combining pictographs between different countries.

ACKNOWLEDGEMENTS

I wish to express my gratitude to everyone in the Institute for Digitization of the Palace Museum Heritage who was enormously helpful with the experiment in this research.

REFERENCES

[1] MSN Messenger: http//messenger.msn.com/

[2] Y. Kawakami, Y.Kawaura, K.Ikeda, R.Hurukawa: *Passport to social psychology computer communications on electronic network*, Seishin Shobo, 1993.

[3] J. Munemori, T. Shigenobu, S. Maruno, H. Ozaki, S. Ohno. T. Yoshino: Effects of Applying Multimedia Video Conferencing System to Intercultural Collaboration, *Journal of IPSJ (Information Processing Society Japan)*, Vol.46, No.1, pages 26-37, 2005

[4] Godai Embody Co.,Ltd. PIC-DIC : http://www.mentek-godai.co.jp

[5] C. Wang: *TOMPA CHARACTERS Living Hieroglyphics*, Maar-sha, 1996.

[6] P. Ardagh: *The hieroglyphs handbook*, Shoeisha, 2000.

[7] NHK the South Pole kids project

http://www.nhk.or.jp/nankyoku-kids/ja/frame.html

[8] NPO Pangaean :

http://www.pangaean.org/common/

[9] The Yomiuri Shimbun:

http://www.yomiuri.co.jp/komachi/reader/200411/2004111500073.html

A Dynamic Control Scheme of Context Information for Context-aware Services

Hideyuki Takahashi[†], Yoshihisa Sato[†,‡], Takuo Suganuma[†,‡], and Norio Shiratori[†,‡]

[†]Research Institute of Electrical Communication, Tohoku University, Japan
[‡]Graduate School of Information Sciences, Tohoku University, Japan
{hideyuki, yosi-q, suganuma, norio}@shiratori.riec.tohoku.ac.jp

Abstract - In ubiquitous computing environments, an effective handling of Quality of Service (QoS) is needed to provide context-aware service in tune with the variation of context information. But QoS for overall system may fall if either the quantity or quality of context information exchanged among entities is insufficient or excess. Based on user context, resource context and QoS, we propose a dynamic control scheme for context information delivery. Based on highly autonomous and cooperative multiagent system, we compose the entities to configure context-aware services. Our proposed scheme can efficiently adapt to the various changes in the real-world environment and maintain the QoS to the user satisfaction. The effectiveness of our proposed scheme is evaluated from the perspective of adaptability depending on the varying relationship between real-word and ubiquitous computing environment.

Keywords: Ubiquitous computing, Context information management, Quality of Service, Quality of Context, Multiagent system

1 INTRODUCTION

In recent years, ubiquitous computing (ubicomp) [1] environment is emerging, where various kinds of sensors, handheld terminals, and wireless networks cooperatively work to support daily lives of the people. As a distinctive service provided by this environment, context-aware services come to the front [2]–[4]. The context-aware services are the services in ubicomp environment based on the "context" which is the situation of each entity. Here, the entity is the element in ubicomp environment such as users, hardware devices, software, and networks.

The context of an entity is delivered in the form of information in the network, and used by other entities. We call this information as "Context Information" (CI). The CI is exchanged at very frequent intervals, then the entities should transfer and handle massive amounts of CI, as well as handling data for main services, using the shared resources. Especially in case of application that consumes a large amount of computational and network resources, such as multimedia communication service, the resources for service provisioning itself decrease due to excessive circulation and processing of CI. Therefore a critical degradation in the QoS in overall system may result.

To solve this problem, Tokairin et al. [10] proposed a Context Management Scheme to keep the QoS. Their scheme is based on the concept of quality of context (QoC), and increases QoS in ubicomp environment by managing CI delivery effectively. However, even if the types of entities are same, the behavior of each entity is different depending on the physical environments where they exist in the real world situation. Hence, the autonomous and adaptive control scheme of CI delivery is needed, according to the varied physical environment in the real world.

Our goal is to develop an effective CI managing scheme for context-aware services to provide appropriate QoS according to varied physical environment and user's requirements. This paper presents a dynamic control scheme of CI delivery based on multiagent technology. To realize this we compose entities which configure context-aware services as highly autonomous and cooperative agents in this scheme. By employing the proposed scheme, the system can provide the context-aware services in varied physical environment flexibly by using the situational adaptability of each agent and their organizational behaviors.

Especially, we performed an initial experiment with the prototype system of a ubiquitous live streaming video service with our proposed scheme. From the results, we confirmed that the system can control the CI delivery according to provided QoS and the change of location and speed of a user entity in the physical environment. We also confirmed the effectiveness of our scheme in perspective of user-level QoS of the video streaming such as frame-rate and timeliness of service provisioning, according to the changes of entities in the real-world environment.

2 RELATED WORK AND PROBLEMS

2.1 Related Work

In this section, we present related works and summarize their problems. Some studies have explored concentrating on the acquisition and selection of context information (CI) for context-aware service provision [5], [6]. CHANSE [5] aims for easy configuration of context-aware service by using centralized management server of CI. In case of failure to get requested CI for configuration of service due to break down of a sensor device, the system can keep CI provisioning by selecting alternative CI from other available sensors. This mechanism gives a good solution to increase availability of context-aware service; however, it is required to describe static process of electing CI when a device is newly introduced. The dynamic reconfiguration of sensor devices is also expected

when the device providing CI breaks down. ContextDistillery [6] proposes a framework that aims for abstraction of the up-to-dateness's variety of each CI. However, it is required to describe the process of selecting CI and up-to-dateness of CI statically in its design phase. The dynamic provision of CI to adapt to the operational situation of the system is expected in the real world.

Additionally, some studies have investigated focusing on Quality of Context (QoC) [7]–[9]. Buchholz et al. advocated the concept of QoC [7]. They introduced and defined from "precision", "probability of correctness", "trust-worthiness", "resolution", and "up-to-dateness" for QoC parameters. They also refer to relation of QoC, QoS, and Quality of Device (QoD). Sheikh et al. [8] define QoC aiming to deal with complicated specification of CI effectively in the middleware for context-aware service. They defined QoC parameters for more practical use than [7], they are, "Temporal Resolution", "Spatial Resolution", and "Probability of Correctness". They mainly discuss the signification of usage of QoC parameter, however, it is not clear how the QoC may be used in system operation for real world applications. An adaptive middleware framework [9] is proposed to provide and select CI depending on the application. They define QoC by "Precision" and "Refresh Rate". This framework computes Utility Function based on the QoC; CI are selected according to the computed amount. This is the pioneering work to use QoC in the real system. However it is difficult to control "Precision" and "Refresh Rate" dynamically because these QoC parameters are assumed to be preliminarily defined and advised.

On the other hand, a flexible QoC control scheme during the system operation is proposed [10]; this study aims to provide the advanced QoS control ability. They propose context information management based on multiagent system; while they focus on "up-to-dateness" as a QoC parameter. This scheme enables the control of CI delivery based on resource status and user's physical location. It is possible to manage QoS of overall system. However, because applicable real environment and services are restricted by the static value of the threshold of resource status and the fixed area of the user's location, which are specified in the design phase of the system, thier scheme has some limitations in scalability and flexibility of QoS control ability.

2.2 Problems

We tackle the control of the context information (CI) delivery. Here, we point out two current technical problems. We assume that ubicomp environment comprises physical environment (real space) and ubiquitous computing space (ubicomp space). We describe the problems of adaptability from viewpoints of relationship between real space and ubicomp space.

(P1) Effective CI delivery based on the run-time behavior of entities: Suppose the cases where context-aware services are provided by using computational and network resources shared by main services and CI delivery in ubicomp space. The resources required to provide its main service is degraded due to the resource limitations of ubicomp space when the amount of CI exchanged among entities becomes huge. Consequently, the main service doesn't work properly or the QoS may be greatly decreased. The effective CI delivery in ubicomp space, deeply considering the run-time behavior of entities in real space, is essential.

(P2) Flexible CI delivery adapted to change of real space in long-term range: We described the problem about reactive adaptability of CI delivery control in run-time in (P1); however, we must deal with the changes of real space in more long-term range. These changes may occur with reorganization of furniture in a room, replacement of the role of a room, modification of work flow, etc. We also need to take the deployment problem of a ubicomp space to a real space into account. It is necessary to customize and adjust the ubicomp space delicately in order to make it workable on the target real space because of diversity of real space. Therefore, it is required to adapt the QoC control scheme autonomously during the service provisioning when a ubicomp space is deployed on an arbitrary real space, and even when the real space changes in long-term range. Then, it is possible for the system to improve its QoC control ability by itself gradually over time.

3 DYNAMIC CONTROL SCHEME OF CONTEXT INFORMATION BASED ON MULTIAGENT

3.1 Relation of QoS and QoC

A general model of context-aware service provision system is shown in Figure 1. The real space comprises four kinds of entities: hardware entities such as PC, PDA, and RFID Tag, software entities such as video transmission/receiving system working on PC, network entities such as wired/wireless networks that connect the entities, and user entities as human user of the system. Here "Context Information" (CI) is defined as the information which carries the situations of these entities.

The entities provide a service for user entities in cooperation with other entities. If the system controls a service based on a specific CI, we call this service "context-aware service." First, the system collects CI from each entity in real space, as shown in Figure 1. "CI Delivery and Handling System" processes the CI and passes only the necessary CI to "Service Provision System" that provides the main service. Service Provision System organizes the entity group to provide the service based on the CI, and the system starts the service for the user entity.

Both the systems in ubicomp space sometimes use the same resources; where resources mean computational resources and network resources. In this situation, if enough amounts of resources are not available, the conflict on the resources occurs and then we cannot get enough performance as described in Section 2.2(P1). Therefore, we need to tune the system to improve the provided QoS as much as possible.

We define QoS and QoC in ubicomp environment as measurement metrics to tune this system effectively. QoS repre-

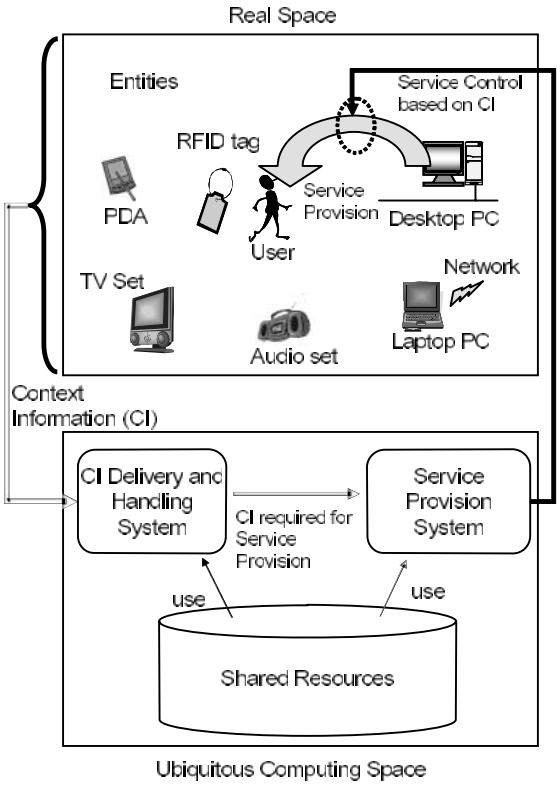


Figure 1: A model of context-aware service provision system in ubiquitous computing environment

sents the quality of service that ubicomp space provides to the users. We define QoS as following three parameters;

$$QoS =< Service{-}Quality, Timeliness, In{-}Placeness >$$

$Service - Quality$ means a quality of providing service itself. $Timeliness$ is an indicator that shows how the service starts in right timing from temporal viewpoint. $In - Placeness$ is a measure that shows how the service is provided in appropriate place from special viewpoint. Second and third items are essential in ubicomp environment, and it is important to provide service that satisfies user requirement to these things.

Additionally, QoC is defined based on [7];

$$QoC =< precision, correctness, trust{-}worthiness, resolution, up{-}to{-}dateness >$$

Here, QoC means the quality of CI exchanged among entities and has tight relationship with QoS. A large amount of resources are consumed when CI is delivered and processed with high QoC. Then the resource that is used by the main service is reduced, and the value of each parameter of QoS decreases in the end; concretely speaking, $Service{-}Quality$ decreases. On the other hands, when QoC is decreased, the context awareness is also degraded. It causes troubles such as service delay in establishment (low $Timeliness$), unexpected starting of service in the place where the users do not exist (low $In{-}placeness$), and so on in these cases. These are fatal errors in context-aware services.

We focus on the relationship between QoS and QoC in this paper. We ensure QoS-aware context service provision by controlling QoC based on the condition of entities in real space and provided QoS, and realize better QoS as much as possible.

3.2 Overview of the Dynamic Control Scheme of Context Information Delivery

We propose an "Dynamic Control Scheme of Context Information Delivery" in this paper. This scheme needs to realizes ubicomp space that can provide effective context-aware services by adapting to various changes in real space. By employing the proposed scheme, the system can improve QoS by managing QoC based on the relationship between QoC and QoS described in Section 3.1. This scheme has the following two functions.

(F1) Dynamic QoC Control Function based on the behavior of entities in real space: This function follows the short-term behavior of entities on the second time scale and automatically adjusts QoC to provide higher QoS and keep it as stable as possible. For example, in the case of ever-changing of the location CI of user entity and hardware entities caused by the user's movement in a room, this function avoids to decrease QoS by increasing/decreasing the QoC of the location CI according to the user's location and movement. This function solves (P1) described in Section 2.2.

(F2) Adaptation Function for long-term changes in real space: This function tunes working parameters of (F1) Dynamic QoC Control Function according to the changing situation of entity for long-term (like in a matter of days) in real space. For example, this function can make the adjustment of algorithm of (F1) adaptively in case the trend of entity's working conditions are changed by the rearrangement of furniture and IT devices inside the room. This can solve (P2) of Section 2.2.

In addition, there needs to be another adjustment for further long-term changes in real space, like the change of number of entities or their quality, occurred by replacement of entity members. In this paper, we omit this type of adjustment.

Compared with existing studies, Tokairin et al. [10] have investigated the QoC adjustment based on user location considering trade-off between QoS and QoC. They suggest a method to vary CI along with the area where the user exists. The areas are determined in advance, and they assign value of QoC per area statically. This gives a partial function of (F1). However, this method lacks of flexibility and extensibility against the changes of real space. Our proposed scheme realizes (F1) fully to apply various real spaces easily. Moreover our scheme can overcome existing method in terms of the adaptability by introducing (F2) which is difficult to realize in existing works.

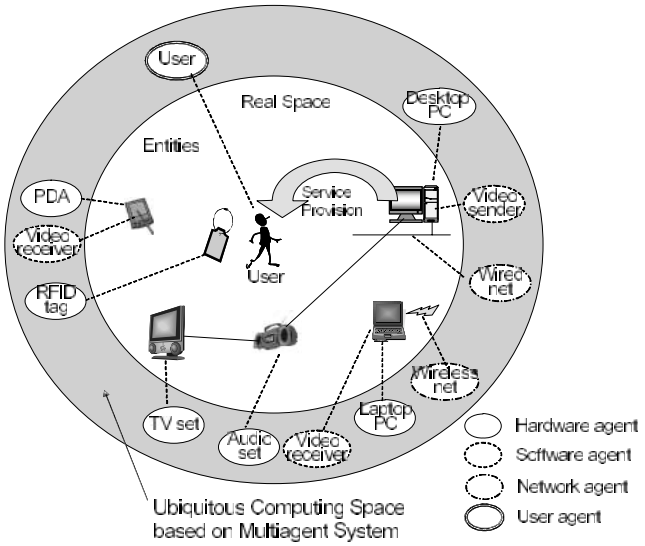


Figure 2: Ubiquitous computing space based on multiagent

3.3 Architecture based on Multiagent

The problems (P1) and (P2) in Section 2.2 originate from the limitations of autonomy and cooperativeness of each entity, and the lack of infrastructure for a flexible system construction to support context-aware services. We therefore compose these entities as highly autonomous and cooperative agents, and construct the ubicomp space as a multiagent system. Figure 2 shows the Ubiquitous computing space based on multiagent. Each entity is monitored and controlled by the individual agent. Each agent manages CI of each entity and exchanges the CI among the agents by using inter-agent communication protocols. The agent also has knowledge on the specification of the target entity. According to this specification knowledge and monitoring results, the agent can recognize the behavioral situation of the entity. The agents can produce an organization based on the contract among the agents, to configure service organization of entities dynamically [11], [12].

We expect the following advantages by using the concept of multiagent.

- Effective service composition in ad hoc manner based on the dynamic selection and synthesis of CI by agent's ability of self-awareness and cooperativeness
- Autonomous acquisition of CI reflecting operational situation of entity by the reflection and autonomy ability of agent
- Advanced provision, delivery and distributed management of CI by the agent's cooperation ability
- Improvement of extensibility and flexibility in the system construction by the modularity and the organizational behavior of agent

3.4 Design of Dynamic Control Scheme of CI

(F1) and (F2) are functions to keep QoS high and stable by monitoring the observable CI that affects to QoS of context-aware service, and by adjusting QoC of the operable CI. In particular, we define (F1) as a mapping function that calculates the value of QoC by using appropriate observable parameter of CI as "variable" and pre-defined "coefficients" of each application. We also define (F2) as a function that updates the coefficients based on evaluation after a single service provisioning is finished. This mapping function is defined for every application and is maintained in the agent that has the decision making role.

In this paper, we consider the design of this scheme by using an application example in which the movement of a user is regarded as the change of situation of the real space. This service can be provided at the right place according to the user's location. Here, we regard the location information of the user entity as the observable CI. We also regard the up-to-dateness of location information of user entity as the QoC of operable CI.

In order to trace the movement of a user correctly, the faster the user moves, the higher up-to-dateness of the location information we need. This is because the system must keep the difference small between the user's actual position and the location information. Furthermore, we need to change the up-to-dateness (QoC) based on not only user's moving speed but also the distance between the user and a service terminal. When the service terminal is far away, we can set low QoC and reduce the consumption of the resource, because the possibility of service provisioning is rather low. We also need high QoC and check the user's location frequently when the user closes to the service terminal, because the possibility of service provisioning increases. By definition of the mapping function that uses the CI as variables, we do not have to set the fixed value manually for mapping between observed CI and QoC. Therefore, we can change QoC continuously based on the changes of real space.

3.5 An Example of Design of Proposed Scheme

We illustrate an example of design of the proposed scheme by using the ubiquitous video streaming service as shown in Figure 3. We consider the following scenario in this service. The user moves in a room while receiving movies with the handheld PC. When the user moves, the system switches the display device and migrates the video player service to a desktop PC that can display the video in higher quality. In this scenario, the desktop PC that can provide the service calculates distance between the user and desktop PC. Moreover it calculates the user's speed from movement distance of the user and an elapsed time from the previous point. The system sends a QoC change request according to the user's speed and the distance between the user and the desktop PC.

Furthermore, QoC should be controlled considering status of QoS to the user. Suppose that the desktop PC shown in Fig-

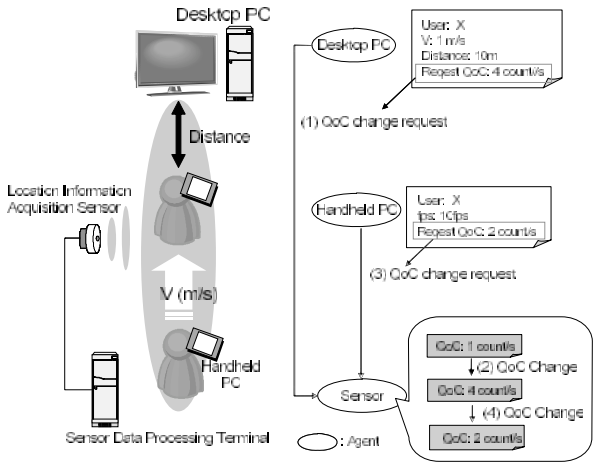


Figure 3: An example of dynamic QoC control function

ure 3, for instance, sets the high QoC, whereas the handheld PC runs out of resources because of the excessive delivery of CI. Thereby it may degrade service quality of video. By monitoring QoS parameters such as frame rate of video, when the actual decrease of QoS occurred, the system requests to change up-to-dateness (QoC) lower. This QoS-based control of QoC can keep the behavior of the entire system stable.

Here is an example of QoC calculation function with a scenario shown in Figure 4. A user moves from point A to point B with handheld PC connected to the Wireless network. The distance between point A and point B are the cover area of the ultrasonic location sensor (ZPS). The service migration occurs when the user get close to the desktop PC at point B. Here, we use the frequency of location information update as the $up{-}to{-}dateness$ parameter of QoC, and frame rate (fps) as the $Service{-}Quality$ parameter of QoS as mentioned in Section 3.1. In this case, we calculate QoC as follow:

$$QoC = \frac{w}{(v+p)\{(X-x')+\alpha\}} \quad (1)$$

where X is the maximum distance between point A and point B as shown in Figure 4, x' is the distance between the user and the terminal that is the destination of the service migration, v is the speed of the user, p is a constant to consider when the user does not move, and w and α are constants to set initial value of QoC when the user stays at point A. The Adaptive Function (F2) is achieved by adjusting each coefficient value X, w, p, and α.

Moreover, the system changes QoC when the average of the frame rate decreases more than that requested by the user. We calculate and update the new QoC value by following simple expression:

$$QoC = \frac{QoC'}{2} \quad (2)$$

where QoC' is the QoC before the request is issued.

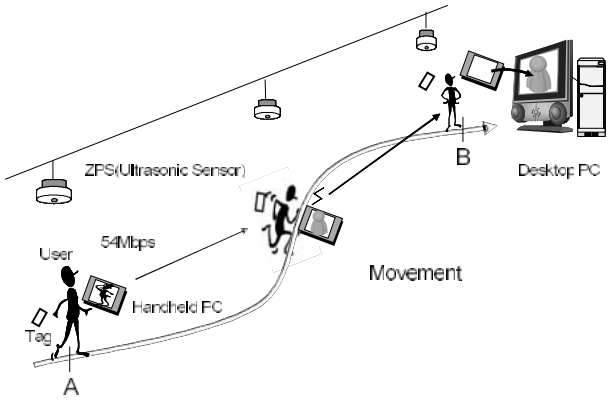


Figure 4: A scenario of QoC control experiment

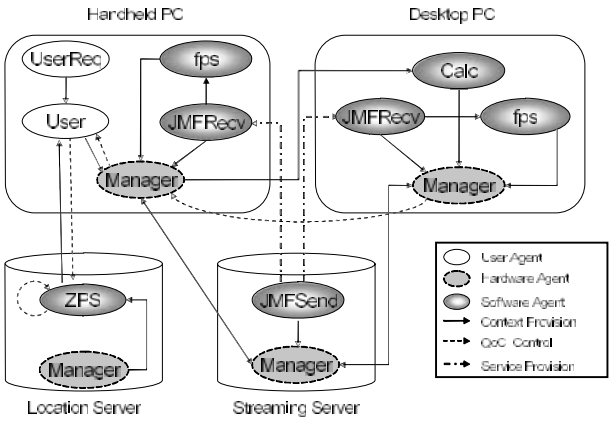


Figure 5: Agent organizations in experimental system

4 IMPLEMENTATION

4.1 An Application to Live Streaming Service for Ubicomp Environments

We developed an experimental system of live streaming service for ubicomp environments. This is a system that transmits a live video captured by multiple cameras to users at remote locations over the network. This system can play the video on a display device such as TV monitor or display of PC which is located at the nearest position to the users. The system has functions to switch the displays and to migrate the video player service according to the user's position and request of QoS. For example, we expect various application areas like a supervising system for the appearance of the elderly person at home, user's baby in the day-care center, and the indoor pet.

4.2 Design of Experimental System

Figure 5 presents the agent's organization used in this experimental system. The role of each agent is shown as follows:

User: This agent manages CI about a user entity. It manages

the user's profile, request of QoS, location information, etc.; it sends the user-related information to other agents.

UserReq: This agent controls U/I software entities that have functions to acquire user request. It informs the user request obtained through this U/I to the *User* agent.

ZPS: This agent has function of acquiring and providing the location information of a specific entity with its tag using an ultrasonic positioning sensor system [14]. This agent can control the up-to-dateness of location information dynamically based on requests (QoC change request) from other agents; this agent informs the CI to other agents.

Manager: This agent manages all the agents on each terminal PC; *Manager* can communicate with the other *Manager*. It judges timing and devices for transmitting and receiving the video service to be switched.

JMFSend, JMFRecv: These agents control Java Media Framework (JMF) [15] which is an entity of multimedia communication software component. *JMFSend* agent transmits streaming video and *JMFRecv* agent receives it based on the request of *Manager*. These agents can adjust the quality of video such as encoding quality, data rate, and frame rate.

Fps: This agent monitors the frame rate of streaming video which *JMFRecv* receives. It passes on a warning to *Manager* when it detects requirement violation about the frame rate.

Calc: This agent calculates the distance between the user and the target PC, and the speed of the moving user using the location information; this agent derives value of QoC by using the mapping function described in Section 3.5.

Figure 6 shows the experimental environment. We employed DASH [11] and IDEA [13] for the software infrastructure. DASH is a rule-based multiagent framework, while IDEA is a development-support environment of the DASH agents. Figure 7 shows a snapshot of a handheld PC and an ultrasonic positioning sensor system (ZPS). We used ZPS as a sensing device for the location information. In this system, a ZPS tag is carried by a user. Thus the location information of the tag is mapped onto the location information of the user entity.

We used C++ and java as program languages of the system build on Windows XP operating system. As for the transport layer, we used TCP in the control part and UDP for the video transmission. We constructed the environment with two kinds of networks: wired (Ethernet 100 Mbps) and wireless (IEEE802.11g, 54 Mbps) networks. In addition, we used a Web camera to capture the video and PC display to receive and play the video delivered from the Web camera.

As for the default setting, a handheld PC which is always carried by a user displays the video form a Web Camera in Figure 6. Moreover, the destination of output of the video migrates from the handheld PC to the display of the desktop PC when the user approaches a desktop PC. *User* and *UserReq* reside in the handheld PC. *User* manages the user's location information receiving from *ZPS*.

5 EXPERIMENTS AND EVALUATION

5.1 Experimental Method

We performed experiments to confirm effectiveness of the proposed scheme. In these experiments, we evaluated that the system can satisfy QoS requirement of user by QoC control which is adapted to the user's location and moving speed. We use a prototype system described in Section 4.2.

We measured frame rate in receiving the streaming video. This corresponds to the $Service{-}Quality$ parameter of QoS described in Section 3.1. We also measured the time that has been spent on the migration of the video streaming from the handheld PC to the desktop PC. This is regarded as the $Timeliness$ parameter of QoS. Moreover, we used frequency of location information update as the $up{-}to{-}dateness$ parameter of QoC.

This experiment uses the scenario of Figure 4. The user moves from point A to point B at a constant speed with the handheld PC. The service is migrated when the user approaches to a desktop PC at point B.

We used Expr. (1) and Expr. (2) which are described in Section 3.5 to determine QoC. In this experiment, the parameters of Expr. (1), they are, X, w, p, and α, were fixed and set to the following values based on the experience of preliminary experiments.

$$X = 5550,\ w = 10^3,\ p = 0.1,\ \alpha = 1000$$

We also assumed that the QoS requirement of the user is 30 fps. In addition, the system considers this situation as QoS violation and sends QoC change request to $Manager$ agent; its request value is decided by Expr. (2) when the 5-seconds average of the frame rate decreases less than 28 fps.

This prototype system can calculate the user's speed (v) dynamically and can use it to decide QoC value. However, to show the effect of the proposed scheme clearly, we set v to the fixed value as 0.1, 0.5, 1.0, and 2.0, and compared the results from each case.

5.2 Experimental Results

Figure 8 shows the experimental results. When the user moved from point A to point B, these graphs represent temporal changes of QoS (fps) of received video and QoC (update

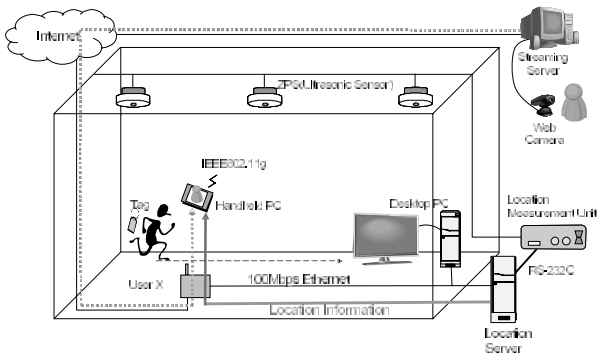


Figure 6: Experimental environment

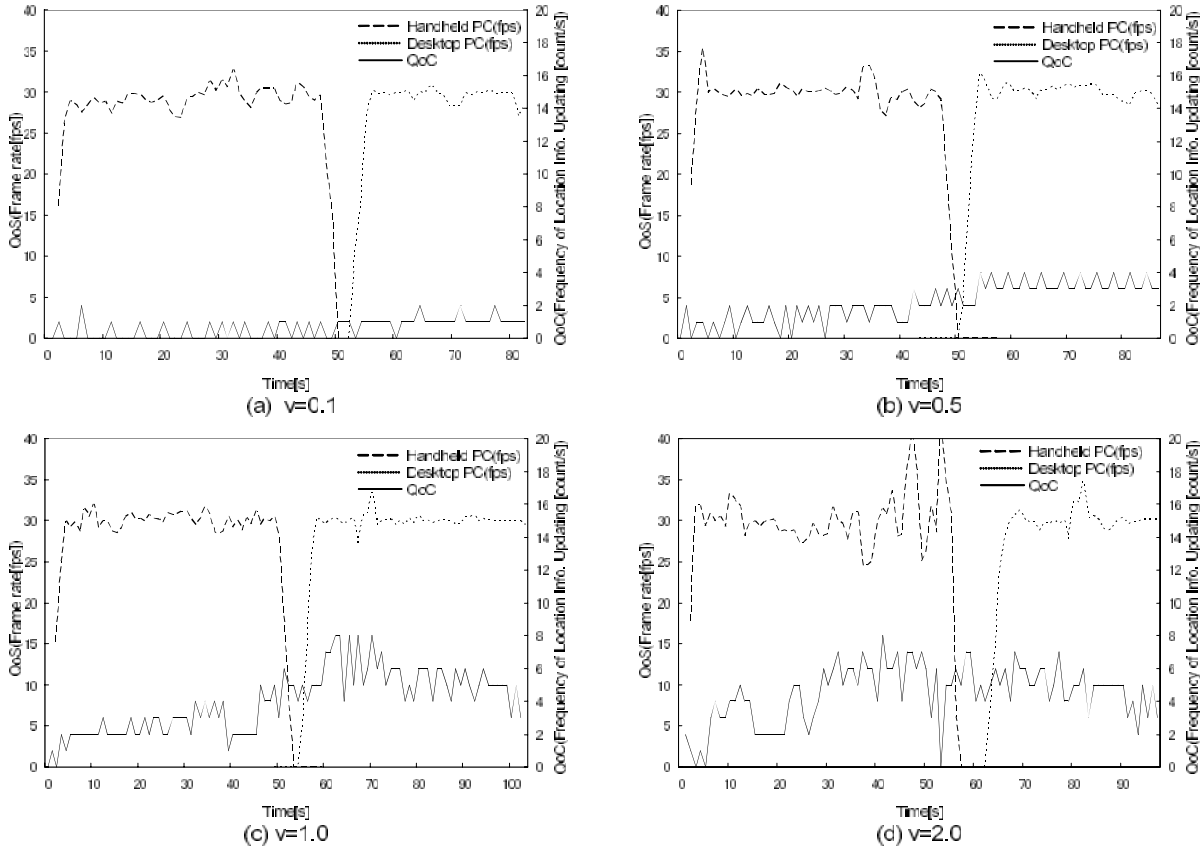


Figure 8: Experimental results of service migration experiment (v as 0.1, 0.5, 1.0, and 2.0)

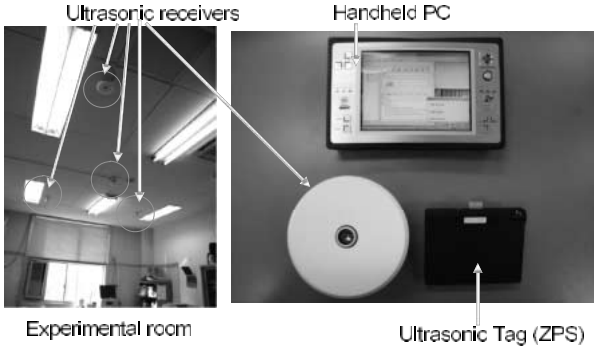


Figure 7: A snapshot of hardware configuration for experimental system

frequency per second). In each graph, fps falls rapidly on the way and afterwards rises again. This shows that the streaming video migrates from handheld PC to desktop PC. In Figure 8, (a)∼(d) are results of each case of fixing the value of v as 0.1, 0.5, 1.0, and 2.0.

In any case, when the user approached point B from point A, QoC was observed to increase. The case of $v = 0.1$ (Figure 8(a)), for example, after the user left from point A at once, QoC was once per 3 or 4 seconds, and then, QoC was updated to once per 2 seconds at about 30 seconds. After that, QoC was updated 1 time/s at about 50 seconds and the service migrated to desktop PC smoothly. In addition, the average frame rate was 29.1 fps and our proposed scheme did not affect to the QoS requirement of the user.

From a viewpoint of difference of moving speed of the user, it was observed that the more the speed increased, the higher QoC became. In case of $v = 0.5$ (Figure 8(b)), for example, after the user left from point A at once, QoC was set 1 time/s, and updated to 2 times/s at about 10 seconds. After that, QoC was updated 3 times/s at about 40 seconds. Compared with the case of $v = 0.1$ (Figure 8(a)), it is found that, the more the user approached the desktop PC, the more the update frequency of QoC increased.

Moreover, in cases of $v = 1.0$ and 2.0, we also observed QoC control behavior based on QoS monitoring. In case of $v = 1.0$ (Figure 8(c)), when 30 seconds passed after the user left from point A, QoC increased to about 3.5 times/s, but when at the point of 40 seconds, it decreased to about 2 times/s, and it increased to about 4.5 times/s again at the 45 seconds point. We can analyze this behavior as follows: fps had decreased at about 35 seconds; subsequently, *Manager* agent on the handheld PC sent the request to decrease QoC to decrease QoS (fps); and then *Manager* agent sent the request to increase QoC when fps was recovered. We can also see

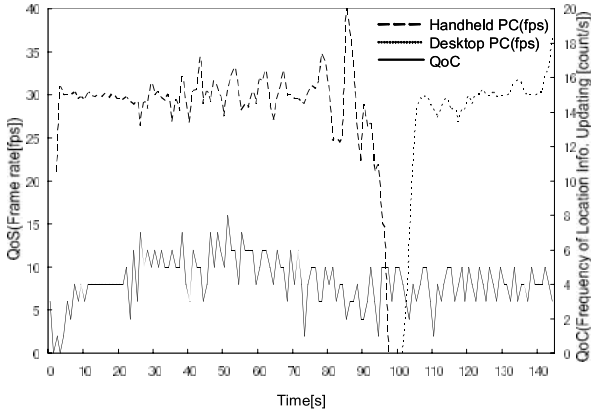


Figure 9: Experimental result in case of no fps agent available ($v = 2.0$)

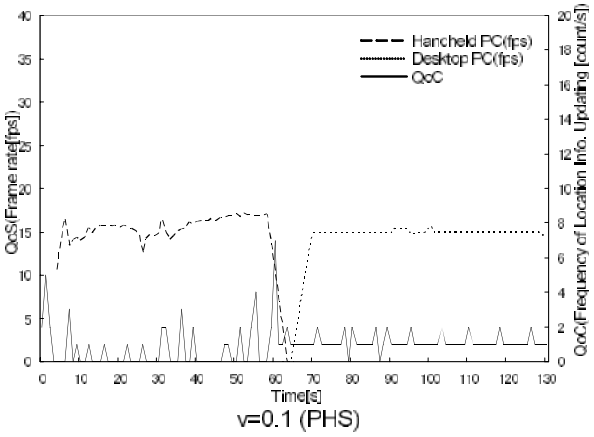


Figure 10: Experimental result in case of using PHS ($v = 0.1$)

the same behavior occurred during 60 seconds to 90 seconds. In case of $v = 2.0$ (Figure 8(d)), the control behavior was observed within the range of 15 seconds to 22 seconds; QoC decreased to 2 times/s. From these results of QoC control based on fps monitoring, the service migration process was successfully executed at about 57 seconds.

We also show a result of the experiment in which the *fps* agent was inactivated. In this case, QoC control based on change of QoS is disabled. The result in case of $v = 2.0$ is shown in Figure 9.

Compared with the situation where fps agent alives (Figure 8(d)), QoC increased gradually and it did not decreased. Moreover, QoC started to decrease from 70 seconds. This is because QoC became 8 times/s at 40 seconds and this excessive QoC caused the unstable frame rate. Therefore, we found that the CPU resource of handheld PC decreased and QoC processing did not catch up.

Additionally we are trying to experiment on using a data communication network of Personal Handy-phone System (PHS) as the network link of the handheld PC. Figure 10 shows the result in case of $v = 0.1$ using PHS/128kbps. In this case, we assumed that the QoS requirement of the user is 15 fps. In the same way as the case of the handheld PC used IEEE802.11g (Figure 8(a)), QoC was observed to increase when the user approach point B from point A, and our proposed scheme did not affect to the QoS requirement (15fps). This result shows that our proposed scheme can apply various network environments.

5.3 Discussion

We show that it is possible to control QoC dynamically based on user's position and moving speed from the experimental results. We also show the problem that providing excessive QoC causes decrease in QoS depending on the user's speed. We confirmed that the system could tune QoC properly to improve QoS by the effect of our proposed QoC control scheme. In this experiment, we did not show the effect of Adaptation Function (F2); however, we give the basis of the function. Further investigation is needed to design and implement the (F2).

Here we discuss the concept of "service session". The service session means a time period when sets of service provisioning are executed in the same configuration of real space. From this classification viewpoint, the function (F1) is regarded as an adaptation in a single service provisioning; whereas (F2) is an adaptation in a single service session based on the evaluation of each single service provisioning. We have to cope with the adaptation through multiple service sessions to realize more effective QoC control. This will contribute to reduce adaptation overhead greatly.

In terms of the improvement of (F1), the existing system keeps same QoC after the user arrives at point B, because QoC control after the service migration is not considered. In the future work, we need to cope with such case when it is unnecessary to increase QoC after service migration; for instance, by suppressing the QoC value.

6 CONCLUSION

We presented a dynamic control scheme of context information delivery based on multiagent, to develop an effective context information managing scheme for context aware service with appropriate QoS according to user's request and changing real space. We also designed and implemented an initial experimental system. From results of experiments by the experimental system, we confirmed the effectiveness of the proposed scheme; it can manage an up-to-dateness (QoC) of positional information according to the QoS, user's position and user's speed. We can greatly improve the adaptation ability of the ubicomp space to the real space by employing our proposed scheme.

We are planning to design and develop (F2) and evaluate it. We only use fps as the QoS parameter and the up-to-dateness of location information as the QoC parameter so far. Our future work includes an extension of the parameter of QoS and QoC. For example, we will use picture size or resolution as a QoS parameter, or accuracy or granularity as a QoC parame-

ter. Moreover we would like to advance detail definition and modeling of relation of QoS and QoC.

Acknowledgement

This work was partially supported by Sendai Intelligent Knowledge Cluster and the Ministry of Education, Culture, Sports, Science and Technology, Grants-in-Aid for Scientific Research, 19200005.

REFERENCES

[1] M. Weiser, The Computer for the Twenty-first Century, Scientific American, Vol. 265, No. 3, pp. 94–104 (1991).

[2] N. Davis, K. Cheverst, K. Mitchell and A. Efrat, Using and determining location in a context-sensitive tour guide, IEEE Comput., Vol. 34, No. 8, pp. 35–41 (2001).

[3] T. Yamazaki, Ubiquitous Home: Real-life Testbed for Home Context-Aware Service, Proc. of First International Conference on Testbeds and Research Infrastructures for the DEvelopment of NeTworks and COMmunities (TRIDENTCOM'05), pp. 54–59 (2005).

[4] E. D. Mynatt, A.-S. Melenhorst, A. D. Fisk and W. A. Rogers, Aware technologies for aging in place: understanding user needs and attitudes, IEEE Pervasive Comput., Vol. 3, No. 2, pp. 36–41 (2004).

[5] T. Nakamura, M. Matsuo and T. Itao, Context handling architecture for adaptive networking services, Proc. of IST Mobile Summit 2000, pp. 295–302 (2000).

[6] K. Fujinami, T. Yamabe and T. Nakajima, Take me with you! : A Case Study of Context-aware Application integrating Cyber and Physical Spaces, Proc. of ACM Symposium on Applied Computing (SAC) 2004, pp. 1607–1614 (2004).

[7] T. Buchholz, A. Kupper and M. Schiffers, Quality of context: what it is and why it need", Proc. of the Workshop of the HP OpenView University Association 2003 (HPOVUA2003) (2003).

[8] K. Sheikh, M. Wegdam and M. van Sinderen, Middleware Support for Quality of Context in Pervasive Context-Aware Systems, Proc of the 5th Annual IEEE International Conference on Pervasive Computing and Communications Workshops (PerComW'07) (2007).

[9] M. C. Huebscher and J. A. McCann, An adaptive middleware framework for context–aware applications, Springer Personal and Ubiquitous Computing Journal, Vol. 10, No. 1, pp. 12–20 (2005).

[10] Y. Tokairin, K. Yamanaka, H. Takahashi, T. Suganuma and N. Shiratori, An Effective QoS Control Scheme for Ubiquitous Services based on Context Information Management, Proc. of International Workshop on NGN and their Impact on E-Commerce and Enterprise Computing(NGN-EC2) in conjunction with CEC'07 and EEE'07, pp. 619–625 (2007).

[11] S. Fujita, H. Hara, K. Sugawara, T. Kinoshita and N. Shiratori, Agent-based Design Model of Adaptive Distributed System, The International Journal of Artificial Intelligence, Neural Networks and Complex Problem-Solving Technologies, Vol. 9, No. 1, pp. 57–70 (1998).

[12] H. Takahashi, Y. Tokairin, T. Suganuma and N. Shiratori, Design and Implementation of An Agent-based middleware for Context-aware Ubiquitous Services, Frontiers in Artificial Intelligence and Applications, New Trends in Software Methodologies, Tools and Techniques (Proc. of the 4th SoMeT2005), Vol. 129, pp. 330–350 (2005).

[13] T. Uchiya, T. Maemura, L. Xiaolu, and T. Kinoshita, Design and Implementation of Interactive Design Environment of Agent System, Proc. of the 20th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE 2007), LNCS4570, pp. 1088–1097 (2007).

[14] Furukawa CO,. LTD., ZPS (Zone Positioning System), http://www.furukawakk.jp/products/ZPS_1.html (in Japanese)

[15] Sun Microsystems, Java Media Framework(JMF), http://java.sun.com/products/java-media/jmf/.

Session 6: Social System
(Chair Teruhisa ICHIKAWA)

Pulse Ejection Presentation System of Odor Synchronized with the User's Breathing

Kaori Ohtsu[†], Junta Sato[†], Yuichi Bannai[‡], and Kenichi Okada[†]

[†]Graduate School of Science and Technology, Keio University, Japan
[‡]Canon inc. Japan
{ohtsu, sato, okada}@mos.ics.keio.ac.jp, bannai.yuichi@canon.co.jp

Abstract - Trials on the transmission of olfactory information together with audio/visual information are currently being conducted in the field of multimedia. However, the continuous emission of odors at high concentrations creates problems of olfactory adaptation and odors lingering in the air which do not match the changes in images over time. To overcome such problems, we applied olfactory pulse ejection to emit odor for very short periods of time. Humans perceive an odor when they breathe in and inhale smell molecules in the air. Therefore, it is important that the timing of pulse ejection is synchronized with breathing. We have developed a breath sensor which detects inspiration, and in this study, we combined the use of this breath sensor with that of an olfactory display in order to establish a pulse ejection presentation system of odor synchronized with breathing. The results of an evaluation experiment showed that the system has a more than 90% detection rate. In addition, a questionnaire survey of the users revealed that the system provides them with a continuous sense of smell, avoiding the effects of adaptation and lingering odor. The use of the developed system is expected to make easier the synchronization of olfactory information transmitted together with audio/visual information.

Keywords: Olfactory information, olfactory display, breath sensor, synchronization with breathing, pulse ejection.

1 INTRODUCTION

Information and communication via computers tends to be limited to visual information and audio information. However, in the real world, humans gather external information via the five senses of sight, hearing, touch, smell and taste, allowing them to react appropriately to local circumstances. Accordingly, the conveyance of such information and its communication via the five senses has lately attracted much attention among researchers in the field of multimedia [1]. Olfactory information recognized by the olfactory organs differs from the information recognized via the other four senses. The sense of smell powerfully affects humans since it is directly transmitted to the cerebral limbic system that governs emotions and memories [2]. Thus, olfactory information serves highly important functions in daily life.

In this study we focus on the development of an ejection control technique for odors, aiming at increasing the application of olfactory information and communication in the field of multimedia. Already in this field, trials are underway to supplement image media such as movies with odors. Any experimental system must control the odor presentation in accordance with the changing images and sounds presented to users. Existing systems have not yet overcome the problem of emitting too much odor over a continuous period, and fine control is necessary in order to avoid various problems such as olfactory adaptation and lingering odors in the air making it difficult to synchronize odors with the ever changing images and sounds.

In efforts to resolve these problems, we attempted to reduce the amount of odor emitted by using pulse ejection for a duration of just 100 msec. In general, it is known that humans can detect odors only when they inhale. So, in order to apply the pulse ejection, it is important that odor presentation is synchronized with breathing [3]. Therefore, we also developed a breath sensor that acquires breath data in real time and can detect the beginning of inspiration. We next combined the developed breath sensor with an olfactory display to create a pulse ejection presentation system of odor that was synchronized with the user's breathing. To examine the performance of this system, we carried out an evaluation experiment and a questionnaire survey with users. This paper presents details of the newly developed system and discusses the evaluation results. [1].

2 RELATED WORK

Trials on the transmission of olfactory information together with audio/visual information are currently being conducted. Work first started in the 1950s when Heilig developed Sensorama [4], the first virtual reality (VR) system that presented olfactory information together with audio/visual information. The recently developed virtual space system, Friend Park [5], provides users with an increased sense of reality by generating the odor of a virtual object or environment. Kaye's article [6] describes some systems that add odor to web content, and computer controlled olfactory displays such as iSmell [7], [8] and Osmooze [9] are utilized in these systems. Another type of display, the air cannon olfactory display that generates toroidal vortices of odor in order to present it in restricted space, has been proposed in [10].

Nakamoto et al. [11] designed a smell synthesis device that presents the odor of a virtual object remotely. The system

[1]This work was supported by SCOPE (the Strategic Information and Communications R&D Promotion Programme) of the Ministry of Internal Affairs and Communications in Japan.

analyzes the smell to be transmitted and presents the analyzed data as the composition ratio of the odor elements. On the receiver side, a feedback control reproduces the target odor by changing the ratio of the odor elements for the receiver.

A wearable olfactory display with a position sensor has also been developed [12]. By controlling the density of odor molecules, it can present the spatiality of olfaction in an outdoor environment. The olfactory information transmitting system consists of the aforementioned display, a sensing system using three gas sensors, and a matching database. The user can experience a real sense of smell through the system by translating obtained olfactory information.

AROMA [13] attempts to introduce the olfactory modality as a potential alternative to the visual and auditory modalities for messaging notifications. Experimental findings indicate that while the olfactory modality was less effective in delivering notifications than the other modalities, it had a less disruptive effect on user engagement in the primary task.

The addition of an odor to image media such as movies has been proposed by a number of researchers. Okada et al. [14] determined the viewer's mental state by measuring his/her brainwaves, and analyzed the relation between the odor and the viewer's feelings while watching. While a movie that supplements visual/audio information with olfactory information has been created, it could not be widely distributed because the synthetic perfume did not correspond with the changing images and the odor was not deodorized.

3 PULSE EJECTION PRESENTATION TECHNIQUE

In this study, we propose an odor presentation technique matched with the individual breathing patterns of the system users receiving the olfactory information. To control olfactory time characteristics, the effects of adaptation and lingering odor in the air must be minimized as far as possible. We therefore use pulse ejection to emit odor for just very short periods of time. Such pulse ejection enables the amount of the emitted odor to be reduced overall, and in a previous experiment, we confirmed that the odor did not remain in the vicinity of the user when presented by pulse ejection with the wind velocity above a certain level [15], thus avoiding olfactory adaption due to scents lingering in the air. Pulse ejection is defined as the olfactory ejection moment that stimulates the olfactory receptors repeatedly and transiently [16], as shown in Figure 1. By using an olfactory display which can provide a stable pulse ejection of 100 msec, we can realize high-precision emission control of odor released into the air.

When humans breathe in, smell molecules in the air are inhaled, and when a smell molecule binds to a receptor organ in the nose, we detect an odor. This is the recognition mechanism of an odor [17]. In addition, air intake in humans decreases during each inspiration [18]; Figure 2 shows we can reliably detect odors only during the early stages of inspiration. The synchronization of odor presentation with the beginning of inspiration is thus considered the most effective technique [19], especially since pulse ejection involves instantaneous odor emission, where the odor disappears almost immediately without remaining in the air.

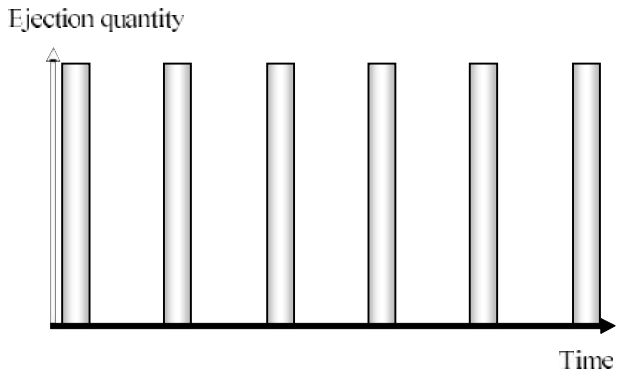


Figure 1: Presentation of bursts of odor via pulse ejection.

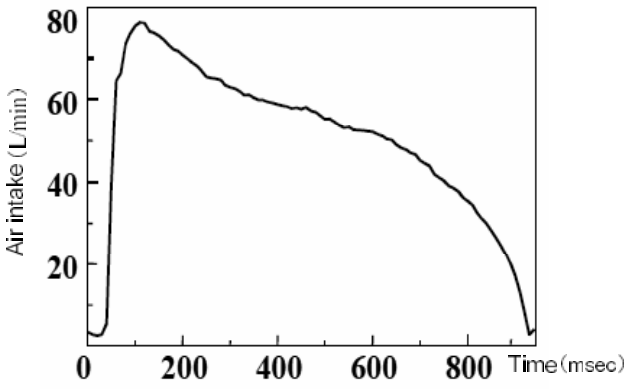


Figure 2: Change in air intake during inspiration over time.

Therefore, to ensure satisfactory recognition of odor, we have developed a pulse ejection presentation system synchronizing odor emission with the user's breathing pattern. This system consists of an olfactory display presenting pulse ejection, a breath sensor acquiring breath data in real time, and a control computer transmitting a signal for scent presentation to the olfactory display. Figure 3 shows a schematic of the developed system.

4 OLFACTORY PRESENTATION SYSTEM

4.1 Olfactory Display

Figure 4 shows the olfactory display developed jointly with Canon Inc. used in the experiment. The inkjet display is able to produce a jet which is broken into droplets by small holes in the ink tank. As the concentration of odor emitted from the display is constant, the display adjusts the perceived strength of the odor by altering the ejection quantity. The display has the following features.

- Twelve kinds of odor tanks

  The display can utilize 3 cassettes, each of which can store one large tank and 3 small tanks, which enables

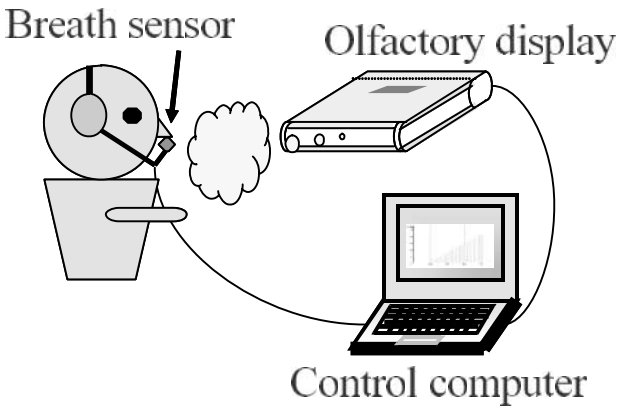


Figure 3: Pulse ejection presentation system synchronized with breathing.

Figure 4: Olfactory display.

the display to present, in total, 12 kinds of odors utilizing 3 large tanks and 9 small tanks.

- Olfactory ejection moment

  Ejection can be controlled for a period of 100 msec.

- Ejection quantity control: 256 phases (large tank), 128 phases (small tank)

  There are 256 minute holes in the cassette connected to the large tank and 128 in the cassette connected to the small tank. The user can control the ejection quantity by changing the number of holes used.

- Wind velocity control: 10 phases

  The display is equipped with a fan that can produce 10 phases of wind velocity control in the range of 0.8 m/sec - 1.8 m/sec.

- Creation of an olfactory scenario

  The user can create a scenario in which olfactory ejection occurs multiple times and can control the amount of odor ejected by altering the tank number, quantity of ejection, ejection start time and ejection end time.

Figure 5: Breath sensor.

For all experiments described below, the display emitted 100 msec pulse ejections of lavender odor. The ejection quantity of the odor was set to 10, which was determined in a preliminary experiment as the value that all users could detect, of the 256 phases. Wind velocity was set to 1.8 m/sec of the display maximum.

4.2 Breath Sensor

To acquire breath information, we developed a breath sensor (Figure 5) which senses temperature change in air inhaled through the nose. The temperature detection element is the NTC (Negative Temperature Coefficient) thermistor which is widely used as a temperature detection element and has a negative temperature characteristic that resistance falls when temperature rises. In this study, we used the NTC thermistor [20] manufactured by Honeywell Inc. An Op-Amp amplifies each item of sensing data, an A/D converter (AMTEC Inc.) converts it to a digital signal, and the value is transferred to a computer.

The data transfer rate of the output voltage level acquired from the breath sensor is 10 sample/sec and the analysis software "TracerDAQ" (AMTEC Inc.) records the data. Figure 6 shows a wave pattern of the recorded breath data from which the beginning of inspiration is detected. Since the temperature of the thermistor falls when air flows during inspiration, the resistance and the output voltage fall. Conversely, the output voltage rises during expiration. Thus, the timing when the wave pattern of breath data begins to fall is judged as the beginning of inspiration.

Characteristics such as breathing intervals differ from person to person, and each user must therefore calibrate the breath sensor before use.

4.3 Pulse Ejection Presentation System of Odor Synchronized with Breathing

Next we developed a pulse ejection presentation system that is synchronized with breathing. The user wearing a breath sensor sits in front of the olfactory display and is presented

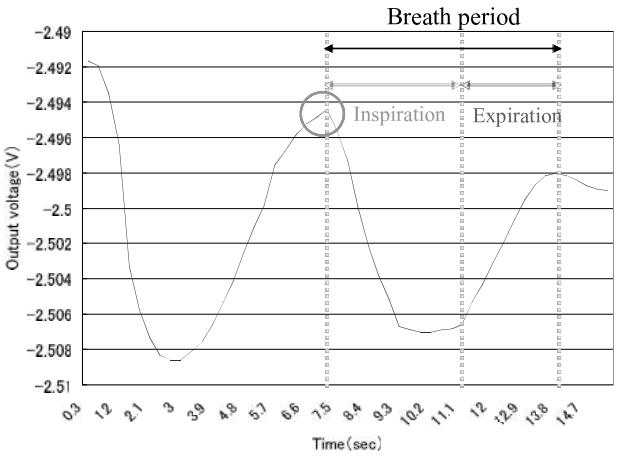


Figure 6: Breath data measured with the breath sensor.

Figure 7: User wearing the breath sensor.

with odor (Figure 7). The system acquires the user's breath data via the breath sensor and transfers the data to a control computer. The control computer runs a program to monitor breath data constantly and to detect the beginning of inspiration. At the point the program judges to be the beginning of inspiration, a signal for odor presentation is transmitted to the olfactory display, which then presents odor to the user. The above represents the process of smell presentation by the pulse ejection presentation system.

5 EXPERIMENT

5.1 System Verification

To verify whether the developed pulse ejection presentation system detected inspiration and presented odor accurately, we conducted a verification experiment with 20 participants (17 males, 3 females, aged in their 20s).

In each experiment, the system monitored around 10 of the participant's breath cycles and presented lavender odor for 100 msec by pulse ejection at the beginning of each inspiration. Participants were instructed to click a mouse when they began to inhale. After the experiment, we verified the performance of the pulse ejection presentation system by comparing the timing of odor presentation as determined by the system with that of the clicking of the mouse button. Each participant performed the experiment two times.

We defined the correct detection rate and the false detection rates as follows and then calculated the rates using data obtained in the verification experiment.

$$Correct\ detection\ rate\ (\%) = NSDC \div NPI \times 100 \quad (1)$$

$$False\ detection\ rate\ (\%) = NSDW \div TNSD \times 100 \quad (2)$$

$NSDC$: $Number\ of\ times\ system\ detected\ inspiration\ correctly$
NPI : $Number\ of\ inspirations$
$NSDW$: $Number\ of\ times\ system\ detected\ inspiration\ wrongly$
$TNSD$: $Total\ number\ of\ times\ system\ detected\ inspiration$

The correct detection rate of this system was determined to be 93.9%, and the false detection rate was 11.3%. As a result, the developed system was confirmed to detect the beginning of inspiration with a probability of more than 90% and to be capable of presenting odor synchronized with breathing. The correct detection rate is able to increase close to 100%, but at present the false detection rate increases with it. Because of this increase in the false detection rate, there is wasteful ejection and an excess quantity of odor is emitted. Depending on the purpose of system usage, it will be necessary to adjust the balance between the correct detection rate and the false detection rate.

5.2 Questionnaire Survey on Feelings Regarding the Odor Presented

We provided 22 participants (16 males, 6 females, aged in their 20s to 60s) an experience of odor presentation using the developed pulse ejection presentation system, and then administered a questionnaire survey in order to determine the users' feelings about the odor presented.

In the experiment, while wearing the breath sensor, participants were exposed to lavender odor that was presented for 100 msec by pulse ejection. The odor was emitted when the system detected the beginning of inspiration and was repeated for 10 breath cycles. As the average time of each breath is 5 seconds [21], it took about one minute to complete the experiment with each participant, and the ejection time of odor was just 1 second in total. After each participant smelled the odor, he/she responded to the following questionnaire items.

Question
What did you notice about odor presented by the pulse ejection presentation system?

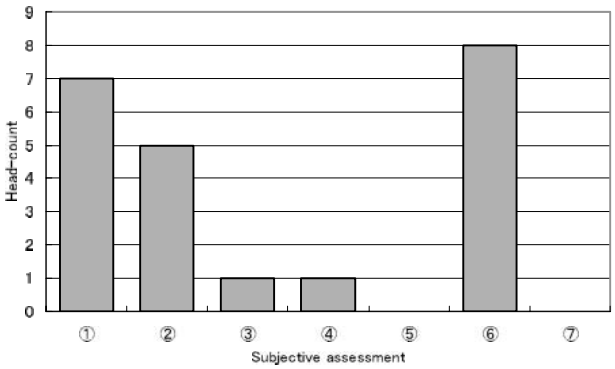


Figure 8: Questionnaire results

The answers presented for the respondent to choose from were as follows:
① Noticed it continuously
② Noticed it on every breath
③ Noticed that the strength of the odor alternated between strong and weak
④ Noticed that the strength of the odor gradually got stronger
⑤ Noticed that the strength of the odor gradually got weaker
⑥ Noticed the odor in fragments
⑦ Did not continually notice the odor

The questionnaire results are shown in Figure 8. Despite the presentation of odor by pulse ejection, many participants noticed the odor continuously (①) or on every breath (②), indicating that the system works effectively. In addition, many participants noticed the odor in fragments (⑥). We think one reason for this is that participants are not able to detect odor during expiration. Another reason is that there were cases when odor was not present because the pulse ejection presentation system likely did not accurately detect inspiration.

It should be noted that there were no participants who noticed the strength of the odor gradually got weaker (⑤) or could not continue perceiving the odor (⑦). This indicates that the presentation of pulse ejection synchronized with breathing could provide the participants with a continuous sense of smell, avoiding the effects of adaptation and lingering odor in the air at least for about one minute.

6 CONCLUSION

In the field of multimedia, trials using odor to supplement audio/visual media are being conducted and it is necessary to control the presentation of odor in order to synchronize it with changes in audio/visual information over time. However, problems of olfactory adaptation and odor remaining in air remain to be solved. To approach these problems, we developed a pulse ejection to present odor synchronized with the inspiration of the receivers of olfactory information.

Pulse ejection of odor for a very short period of 100 msec was stably presented during inspiration by combining the use of a breath sensor which could acquire breath data in real time with an olfactory display that has high emission control.

System verification experiments showed that the system could detect the beginning of inspiration with a probability of more than 90% and present odor synchronized with breathing. In addition, a questionnaire survey of users' opinion of the odor presented by the system revealed that most users could notice odor continuously or on every breath during the trials and none were considered to have been affected by adaptation and lingering odor in air.

The developed system will make the fine control of odor presentation possible, enabling further advances to be made in the transmission of olfactory information together with audio/visual information. As a result, the synchronization between media is expected to become easier.

REFERENCES

[1] Jeong-Do Kim, Dong-Jin Kim, Dong-Won Han, Hyung-Gi Byun, Yu-Kyung Ham, Woo-Suk Jung, Jun-Seok Park and Sam-Kweon Oh, A Proposal Representation, Digital Coding and Clustering of Odor Information, Computational Intelligence and Security, 2006 International Conference, Vol. 1, pp. 872-877 (2006).

[2] G.A. Michael, L. Acquot, J.-L. Millot and G. Brand, Ambient Odors Modulate Visual Attentional Capture, Neuroscience Letters 352, pp. 221-225 (2003).

[3] T. Tonoike, Recording and Analysis of Olfactory Evoked Potentials on the Human Scalp, Summaries of Reports of the Electrotechnical Laboratory 863 (in Japanese), pp. 1–76 (1968).

[4] Retrofuture: Sensorama's pre-virtual reality, http://www.retrofuture.com/sensorama.html

[5] H. Shigeno, S. Honda, T. Osawa, Y. Nagano, K. Okada and Y. Matsushita, A Virtual Space Expressed the Scent and Wind -A Virtual Space System "Friend Park", Journal of Information Processing Society of Japan (in Japanese), Vol. 42, No. 7, pp. 1922–1932 (2001).

[6] J. Kaye, Making Scents: Aromatic Output for HCI, Interactions, Vol. 11, No. 1, pp. 48–61 (2004).

[7] D. A. Washburn, L. M. Jones, R. V. Satya, C. A. Bowers and A. Cortes, Olfactory Use in Virtual Environment Training, Modeling and Simulation Magazine, 2, No. 3 (2004). http://www.modelingandsimulation.org/issue7/olfactory.html

[8] Edge Review: DigiScent Ismell, http://www.edgereview.com/ataglance.cfm?category=Edge&ID=136

[9] http://www.osmooze.com/osmooze/osmooshop_gb.html

[10] Y. Yanagida, H. Noma, N. Tetsutani and A. Tomono, An Unencumbering, Localized Olfactory Display, CHI '03 Extended abstracts, pp. 988–989 (2003).

[11] T. Nakamoto, Y. Nakahira, H. Hiramatsu and T. Morizumi Odor Recorder Using Active Odor Sensing System, Sensors and Actuators B, 76, pp. 465–469 (2001).

[12] S. Yokoyama, T. Tanikawa, K. Hirota and M. Hirose, Olfactory Field Simulation Using Wearable Olfactory

Display, Trans. of Virtual Reality Society of Japan (in Japanese), Vol. 9, No. 3, pp. 265–274 (2004).
http://www.cyber.rcast.u-tokyo.ac.jp/project/nioi_e

[13] A. Bodnar, R. orbett and D. ekrasovski, AROMA: Ambient Awareness through Olfaction in a Messaging Application, ICMI '04 Proceedings, pp. 183–190 (2004).

[14] K. Okada and S. Aiba, Toward the Actualization of Broadcasting Service with Smell Information, Institute of Image information and Television Engineering of Japan Technical Report (in Japanese), Vol. 27, No. 64, pp. 31–34 (2003).

[15] A. Kadowaki, J. Sato, Y. Bannai and K. Okada, Measurement and Modeling of Olfactory Responses to Pulse Ejection of Odors, Japan Association on Odor Environment (in Japanese), Vol. 39, No. 1, pp. 36–43 (2008).

[16] A. Kadowaki, J. Sato, Y. Bannai and K. Okada, Presentation Technique of Scent to Avoid Olfactory Adaptation, 17th International Conference on Artificial Reality and Telexistence, pp. 97–104 (2007).

[17] Bettina Malnic, Junzo Hirono, Takaaki Sato and Linda B. Buck, Combinatorial Receptor Codes for Odors, Cell, Vol. 96, pp. 713–723 (1999).

[18] T. Honma and H. Wakamatsu, Telemedicine by the Compact Type of Assist-Respirator with a New Ventilation Pattern (in Japanese), The Japanese Society of Clinical Physiology, Vol. 31, No. 6, pp. 297–304 (2001).

[19] M. Wada, An Approach to the Study of Brain Waves Induced by Olfactory Stimulation, Aroma Science Series 21, Essay on the Study of Scents (in Japanese), Fragrance Journal Ltd. (2005).

[20] RS Components Ltd,
http://www.rswww.co.jp/cgibin/bv/home/Home.jsp?cacheID=jpie

[21] T. Shibuya and T. Tonoike, Aroma Science Series 21, Odor Receptors (in Japanese), Fragrance Journal Ltd. (2002).

Design of User-oriented Healthcare Support System based on Multi-agent

Hideyuki Takahashi[†], Satoru Izumi[†,‡], Takuo Suganuma[†,‡], Tetsuo Kinoshira[‡,††], and Norio Shiratori[†,‡]

[†]Research Institute of Electrical Communication, Tohoku University, Japan
[‡]Graduate School of Information Sciences, Tohoku University, Japan
[††]Cyberscience Center, Tohoku University, Japan
{hideyuki, izumi, suganuma, norio}@shiratori.riec.tohoku.ac.jp, kino@riec.tohoku.ac.jp

Abstract - In this paper we propose an advanced healthcare support system in ubiquitous computing environment. By utilizing knowledge about healthcare and various information including vital sign, physical location, and video data of a user under observation from real space, the system provides useful information regarding health condition effectively and in user-oriented manner. In this paper, we describe a user-oriented healthcare support system based on concept of symbiotic computing, focusing on design and implementation of the system with multi-agent system.

Keywords: Healthcare support system, Ubiquitous computing, Multi-agent system, Context-aware service, Multimedia communication

1 INTRODUCTION

With the increase of people with lifestyle-related diseases such as obesity, hypertension, diabetes, and hyperlipidemia, health maintenance to prevent these diseases has been an issue of social concern. Information technologies are expected to give practical solutions to this issue, and some research groups have been investigating the solutions from engineering viewpoints [1]–[7]. In this context, ubiquitous computing technologies are promising, because they contribute to expand the scope of system support to users' daily lives. Handheld terminals, wearable vital sensors, wireless communications, etc. are playing important roles in this application domain [8]–[18].

However, these existing systems are designed by using some specific vital sensors and electronic devices, therefore these systems are limited in ability of healthcare support. In order to provide useful information for healthcare of an object person, not only to him/herself but also to related people of the person, the system should acquire variety of information, knowledge, data, etc. from real space and store/manage them in a methodical manner. This means that we have to treat a new dimension of design and construction of large-scale systems that can cope with many kinds and amount of information on unstable processing environment of ubiquitous computing.

We have been investigating an advanced healthcare support system in ubiquitous computing environment. By utilizing knowledge about healthcare and various kinds of information obtained from real space, the system provides useful information regarding health condition effectively and in user-oriented manner. In this paper, we describe a design of user-oriented healthcare support system based on symbiotic computing [19]–[21], that is a concept of post-ubiquitous computing according to co-existing of real-space and digital-space. Especially, we focus on the design and implementation of the system with multi-agent technology that matches to realize this kind of large-scale and complex systems by employing such properties as autonomy, cooperativeness, and adaptability of agents. We also show the effectiveness of our prototype system with results of initial experiments.

2 RELATED WORKS AND PROBLEMS

2.1 Related Works

In this section, we present related works about healthcare support systems, and summarize their problems.

Administrative organizations provide various kinds of information about healthcare on the Web [1], [2]. Several companies have developed a medical device and provide a healthcare service utilizing the device [3].

Some research groups developed support systems which recognize health condition of a user by monitoring user's vital signs using compact sensors, hand-held PCs, and wireless network in ubiquitous computing environment [8]–[13]. There exists the system which can infer user's behavior, activity, and emergency situation according to the vital signs and location information of the user's by using wearable sensors. In [14], Chang et al. studied methods that automatically recognize what type of exercise the user is doing and how many repetitions he/she has done so far. They incorporated a three-axis accelerometer into a workout glove to track hand movements and put another accelerometer on the user's waist to track body posture.

A project [5] was promoted to develop a prototype of next generation network system which can provide high quality health service with sufficient security and protection of user's privacy. Under the project, a health advice derivation system has been developed [6]. This system can derive health advice according to the user's condition and knowledge about health.

2.2 Problems

From discussion on previous works in Section 2.1, we point out technical problems in existing healthcare support system as follows.

(P1) Effective acquisition of various and amount of information related to healthcare in real space

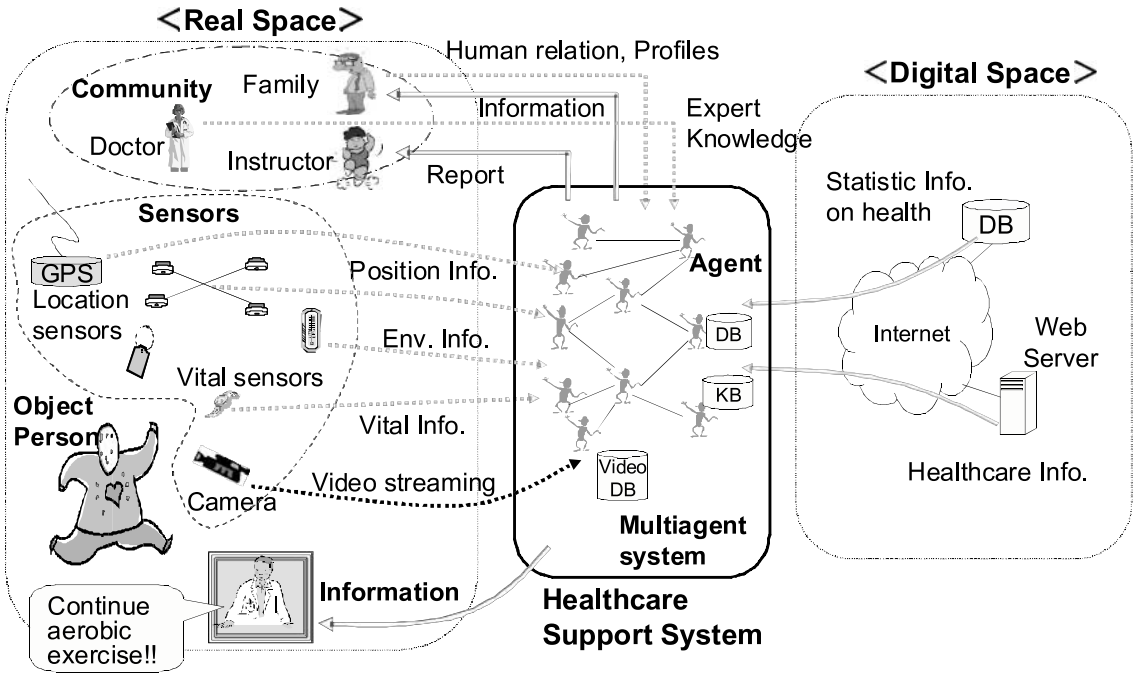


Figure 1: Concept of user-oriented healthcare support system based on multi-agent

There are studies determining the health condition based on vital sign by specific sensing devices in real time. But the information has limitations for obtaining an accurate estimation of the health condition because the information is obtained by the vital sign limited piece of information on certain individuals. It would be possible to perceive the health condition of object person with greater accuracy using physical location of the person, environmental information such as ambient temperature and room brightness, and video information of the person, as well as the vital sign. However, it is difficult to acquire all the information in real space because of the limitation of computational resources and network resources in the ubiquitous computing environment. Consequently, we need to consider the effective way of information acquisition from real space.

(P2) Service provision based on various kinds of information of real space

After acquisition of various kinds of information from real space, effective information and service provisioning using the information would be a challenge. The data and information including vital sign, location information, environmental information, multimedia data, specialized knowledge, etc. contain significant diverse aspects in both quantitative and qualitative. Therefore, it is essential that the mechanism should be given for actively-provisioning of required data and information to a particular user who really needs, in appropriate place, with suitable format, only when the user needs, by effective management.

(P3) Infrastructure of system construction

In the related works, specialized systems in each area of healthcare have been developed in an ad-hoc manner. Thus, we do not have an infrastructure of system construction to facilitate implementation of systems for various healthcare areas. The infrastructure needs system extendibility to introduce new sensor device, wireless network technology, algorithm for analyzing, etc. in easy ways. In order to reduce development costs and facilitate effective implementation of the system, we need common software infrastructure comprised of platform and components dedicated to healthcare support.

3 CONCEPT OF USER-ORIENTED HEALTHCARE SUPPORT SYSTEM BASED ON MULTI-AGENT

3.1 Overview of User-oriented Healthcare Support System

We propose a methodology of construction of user-oriented healthcare support system based on multi-agent to solve the problems mentioned above in Section 2.2. Figure 1 shows the concept of our proposed system.

This system supports a healthcare of an object person and community members who are related to the object person such as family member, sports gym instructor, doctor, etc. to circulate healthcare related information and knowledge effectively. The system collects information on the object person such as profiles, preferences, history of exercise, medical records, human relation, etc. from the healthcare community members. The system actively observes the current status of the object person and his/her surrounding environment such as physical location, temperature, body warmth, HR, BP, etc. by using various types of sensors. These are the information flows from real space to this system.

On the other hands, the system accesses to the Web site and

databases (DBs) via the network to fetch useful information on healthcare. These are the information flows from digital space to this system. The information, data, and knowledge are accumulated in the system in adequate forms. If needed, they are used to analyze the situation of the object person in detail. The information is sometimes provided to the person and the community members by proper timing and forms, considering privacy concerns and resource limitations of the devices.

From the viewpoint of the symbiosis between real-space and digital-space, this support system is an accelerator of information circulation in order to promote the healthcare tasks. However, huge amount and functional diversity of the information, involving the privacy concerns, make it very difficult to accomplish.

3.2 Applying Multi-agent Technology

Multi-agent system is a distributed autonomous coordination system. Various types of system component are wrapped (this wrapping is called "agentification"), and then it gets possible to work as an agent. The multiple agents can dynamically configure organization to process some intended tasks.

Consider the situation where some vital data or location information is acquired by a sensor device, transmitted via the network and stored in a DB. Each agent individually resides in various sensor devices. The agent monitors and controls corresponding hardware. Also the DB which stores acquired data is made to work as agent. Quality and frequency of the acquired data should be controlled depending on network status, operational condition of the sensor device, and load of the DB. The proposed system can effectively control the data flows by cooperation among sensor agent, DB agent, and network agent. This will be a solution to (P1).

The accumulated information is basically in the form of raw data. It should be converted into more user-friendly forms such as tables and graphs. Some data can be used to analyze the situation of the object person to create knowledge or advice with high-level expression. These analytical results can be used by agents' organizational behaviors. For example, when the object person is in bad health condition, the sensor agent that observes the vital data of the person would try to acquire more detailed information in shorter time intervals. To realize these kinds of intelligent analysis, each agent has basic inference mechanism based on the rule-base system. For more special knowledge processing, some kinds of powerful tools, such as ontology-base, data mining algorithms, software for statistics, etc. can be incorporated by the agentification. This would be a solution to (P2).

In addition, agentification of various devices, database, knowledge, algorithm for analysis, software components, etc. makes reusable module, and agents can dynamically configure a complex system. It is possible to build a new component into the existing system at the lowest possible cost when the component is introduced. Thus the infrastructure of system construction base on multi-agent system will realize reduction of system development cost and advancement of the system. This is another important aspect for ubiquitous applications whose technologies are proceeding at a rapid rate daily. This will resolve the (P3).

4 DESIGN OF USER-ORIENTED HEALTHCARE SUPPORT SYSTEM

4.1 Agent-based Framework AMUSE

We employ a multi-agent-based framework for service provisioning in ubiquitous computing environments based on concept of symbiotic computing [19]–[21], called AMUSE (Agent-based Middleware for Ubiquitous Service Environment) [22], [23], as a software platform to build user-oriented healthcare support system. The fundamental framework of AMUSE is shown in Figure 2. The basic idea of this framework is "agentification" of all the entities in the ubiquitous computing environments. Here, the agentification is a process to make a target entity workable and context-aware. The details are discussed in [22], [23], so we omit them in this paper.

In AMUSE, agents can perform the following advanced cooperation and intelligent behavior for the ubiquitous computing:

(1) Recognition of statuses of individual entities

(2) Coordination of contexts of multiple entities

(3) Service construction by combinations of entities

(4) Intelligent processing of the situations around the entities

4.2 AMUSE Agents for Healthcare Support System

We show one example of healthcare support system. We suppose the healthcare support system consists of various daily life support systems such as health management system, multimedia supervisory system, remote medical care system, etc. In this paper, we assume a multimedia supervisory system for healthcare support system. The multimedia supervisory systems are widespread as care-support systems that enable supervision of children and elderly people from remote sites connected by a wide-area network. Figure 3 shows organization of agents in real-time multimedia supervisory system that delivers live video streaming captured with cameras at the watched person's site, with a PC or a hand-held device at the distant supervisor site.

The agents consist of Device-aware agents, User-aware agents, and Social-aware agents based on symbiotic computing. The agents basically reside in computers, and they manage corresponding entities that are connected to, or are running on the computer.

The Device-aware and User-aware agents cooperatively work to accomplish QoS that meets to user's requirements on a watching task and device situations. Multiple contexts of the devices should be deeply considered to achieve our goal. These contexts are individually maintained by each agent, and

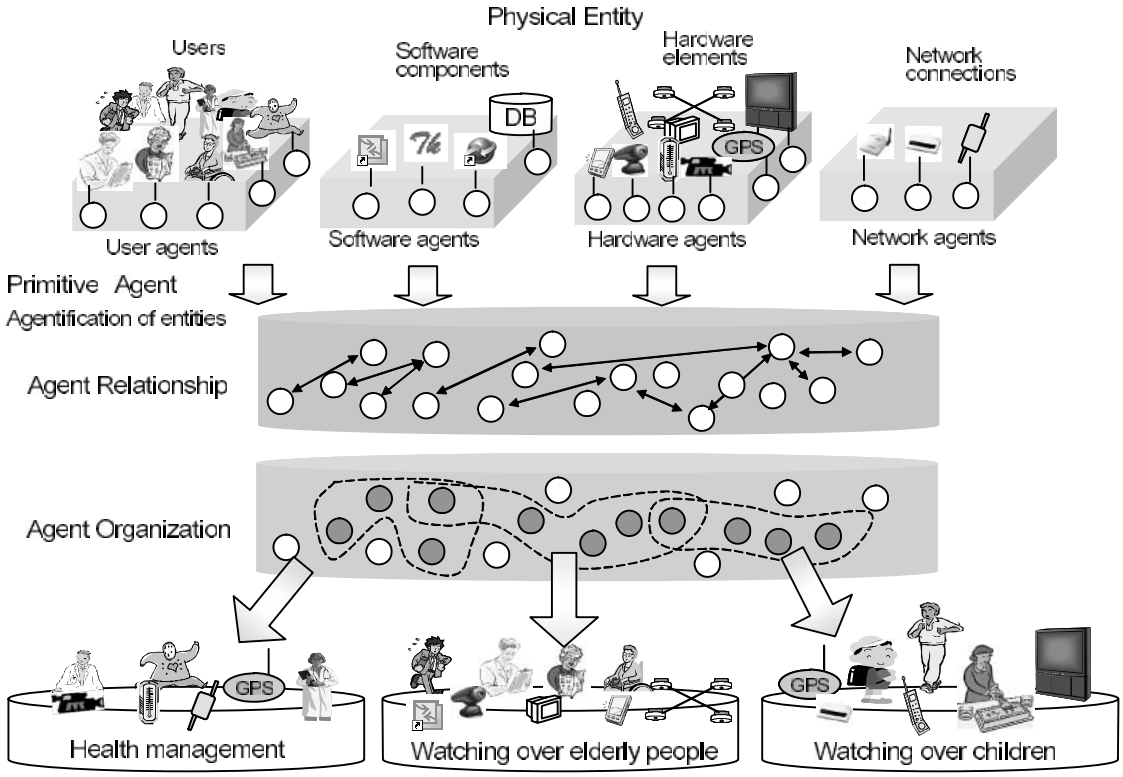


Figure 2: Framework of AMUSE

its effective coordination would be performed by cooperation among related agents.

We summarize the agent library for AMUSE Framework as shown in Table 1. The agents are categorized into four classes: hardware, software, network, and user agents. The base process is the corresponding entity for each agent. We constructed these agents using ADIPS/DASH [24], [25].

5 IMPLEMENTATION

We are developing an application to watch over elderly people. Agents were implemented based on AMUSE framework. As for implementation of agents, we used DASH [25]. DASH is an agent-based programming environment based on ADIPS [24]. We also used IDEA [26] for the development and simulation of the agents. It is an integrated development tool for the DASH.

To evaluate feasibility and effectiveness of our system, we introduce following application scenarios. We suppose a situation where a son watches over his farther (an elderly person) in remote sites. Figure 4 shows three experimental rooms.

Figure 4(a) and Figure 4(b) are regarded as the son's office and his living room in his home. They are watching sites. Additionally, Figure 4(a) and Figure 4(b) show the room settings of the watching sites. He moves around the office room (a) and his living room (b). Here, "Small PC12" represents the watcher's user terminal which is shown in Figure 4(d). This terminal is always brought with the observer. This terminal is selected for receiving the video of the watched person, when other displays cannot be available. A User agent resides in this terminal. The agent monitors the user's requirements and presence. Also, we used ZPS ultra-sonic sensor [27] to sense the watched person's location information in both of the room (a) (Figure 4(a)) and the room (b) (Figure 4(b)).

Figure 4(c) is supposed to be a living room in his father's home (a watched person's home). This is a watched site. Figure 4(c) shows a room setting of the watched site. As for location sensor, we employed an active-type RFID system [28] in this room (c).

Our system displays a live video with suitable quality on one of the displays considering the son's requirement for the watching over and the status of devices. Therefore, our system makes the construction of agent organization by considering the most appropriate camera, the PC with reasonable network connection, and the display devices based on multiple contexts.

6 EXPERIMENTS

6.1 Experiments based on user requirement

In anticipation of our healthcare system, we are developing part of the real-time multimedia supervisory function that delivers live video streaming. This section describes some examples of behavior of the function.

In this experiment, we observe our system behavior based on user requirement. A watching person specifies a user requirement from "best resolution" or "best smoothness" op-

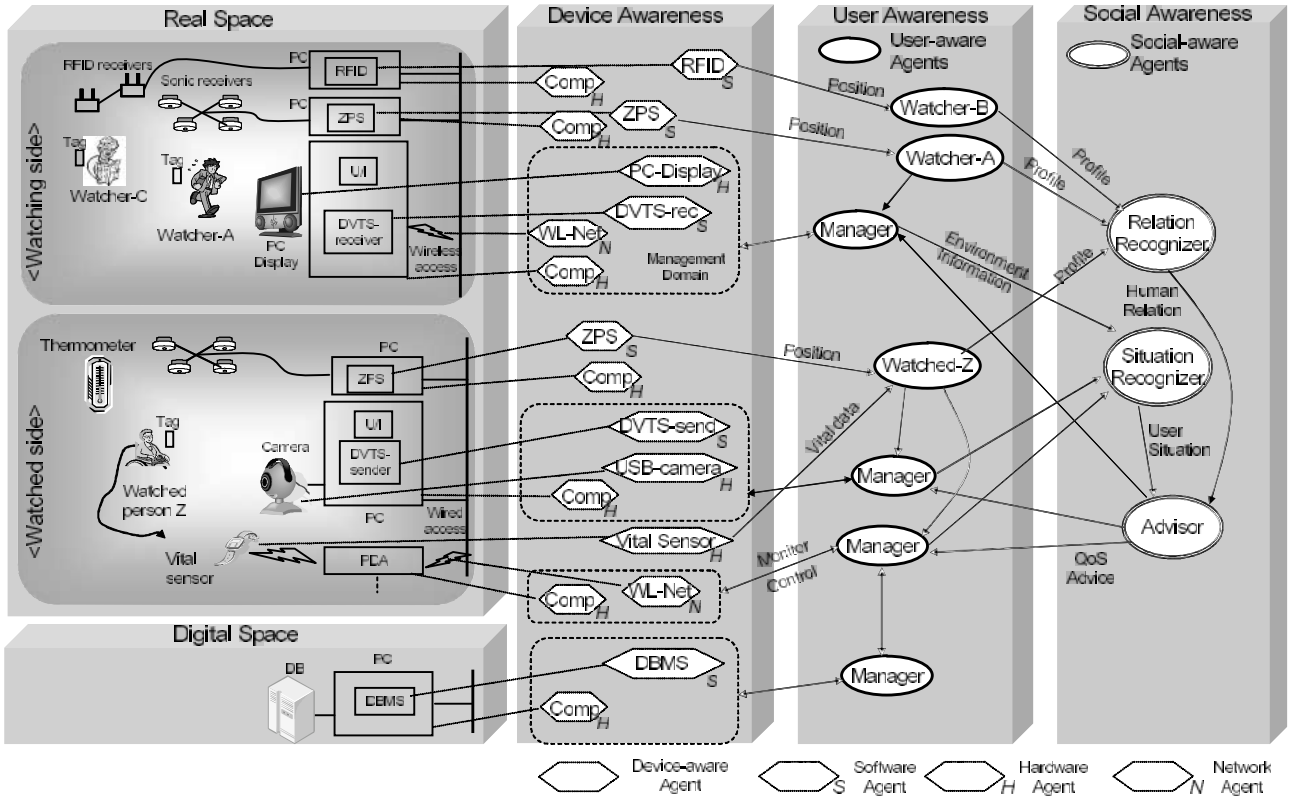


Figure 3: Organization of agents in healthcare support system

tions. The User Ag provides a user interface about the option on the user's terminal. This decision of the option is based on the background of followings:

(1) Best resolution: The son wants to watch the father's facial color or expression in high resolution of the video, because he is worrying about the status of his father's sickness.

(2) Best smoothness: The son wants to see the full-length of his father as smooth as possible, because he cares for his father's health condition.

We compare our system with a scheme to make the effectiveness of our system. We assume the scheme as a location-based service configuration. This scheme selects the nearest camera and display to the watched/watching people, respectively, without any consideration of total quality of the service. We test this configuration using our system with terminating some agents that provide multiple contexts.

The son moves in the rooms (Figure 4(a) and Figure 4(b)). Additionally, we fix his father's location for simplification at point "A" in Figure 4(c). Agents cooperatively work together to select the most adequate sets of entities based on the son's requirements and location.

Case of best resolution: First, the son specifies "best resolution" of the video to watch his father's facial color in this case. When the son moves to the location at point "B" in Figure 4(a), a PC display of PC6 and a TV connected to PC5 can display a live video. It means the point "B" is the video service area for the PC display and the TV.

Here, in case of the location-based scheme, because the PC display of PC6 was judged as the most nearest display, the video service migrated to the PC display of PC6 from the user terminal, as shown in Figure 5(a). However, the quality of the video is too low to see the father's facial color vividly, because it moved with just the same video quality parameters as it was in the user terminal.

On the other hand, in case of our proposal-based scheme, the TV display and the DV camera was selected to fulfill the user's requirement of high resolution as shown in Figure 5(b). In terms of the software entity, our scheme selected DVTS [29] because it can provide the high quality of video. In this case, we confirmed that our scheme could effectively coordinate the multiple contexts, and our scheme could satisfy user requirement.

Case of best smoothness: Next, the son specifies the high smoothness of movement of the video to watch in his father's health condition. When he moves to the location at point "D" in Figure 4(b), a portable PC10 and a PC display of PC11 can display the video. It means the point "D" is the service area of the portable PC and the PC display.

In the case of a location-based scheme, because the display of PC10 was judged as the nearest PC display, the video service moved to display of PC10 from the user terminal, as shown in Figure 6(a). However, the video frame rate was too low to view the movement of his father's body smoothly because it was moved with the same video frame rate parameters

Table 1: Description of agent library

	Function	Agent Name	Base Process	Description
Hardware	Location information	ZPS	ZPS	Identifying location of tags using ultrasonic sensor
		RFID	RFID System	Identifying location of tags using RFID system
	Image input	DV-camera	DV camera	Control a DV camera to capture video image.
		USB-camera	USB camera	Control a USB camera to capture video image.
	Image output	PC-Display	PC Display	Control a connected PC display to show video image.
		TV	TV set	Control a connected TV set to show video image.
	Audio input	Mic	Microphone	Control a connected microphone to capture audio.
	Audio output	Speaker	Speaker	Control a connected speaker to play audio.
	Computer monitoring	Comp	CPUcheck	Monitoring of the status of computational resources such as CPU usage rate in a target computer.
	Biological information	Bio	Heartbeat	Monitoring target person's heartbeat.
Software	Location information	Location_manager		Management of the up-to-dateness specified by the other agent or application developer.
	Video receiver	DVTS-rec	DVTS application	Video receiving in very high quality by using DVTS Software.
		JMF-rec	JMF application	Video receiving in various formats by using the Java Media Framework (JMF).
	Video sender	DVTS-send	DVTS application	Video sending in very high quality by using DVTS Software.
		JMF-send	JMF application	Video sending in various formats by using the JMF Software.
	Management	Manager		Management of behavior of all the agents in the corresponding PC
	Interface	UserReq	U/I component	Maintenance of the GUI-based software component to acquire the user request directly.
	Human relation	Human-Relation-Ontology	Ontology base	Management of the knowledge on human relationship of users.
	Daily activity support	Daily-Activity-Ontology	Ontology base	Maintenance of knowledge on daily activities of users.
	Common sense support	Common-Sense-Knowledge	Knowledge base	Maintenance of common knowledge used in the target application.
	Situation recognition	Situation-Recognizer		Recognition of situation of a target user.
	Relation recognition	Relation-Recognizer		Specifying human relationship between users.
	Decision making	Advisor		Making decisions of action for a specific application.
	Database management	DBMS	DBMS	Management of the data such as vital sign, environmental information, and location information, and multimedia data.
	Technical knowledge support	Technical-Knowledge	Knowledge base	Management of the knowledge on experts for healthcare.
Network	Network monitoring	W-Net	NETcheck	Monitoring status of an wired network.
		WL-Net	NETcheck	Monitoring status of an wireless network.
User	User manager	User name		Management of requirement, preference, profile, etc. of a user.

as it was in the user terminal.

At the same time, Figure 6(b) shows the behavior of our proposal-based scheme. Our scheme selected the PC display of PC11 and the USB camera connected to PC1, with high frame rate to fulfill the user's requirement. As for the network context, PC11 is the best because it is connected by a wired link with 100 Mbps. Additionally, agents recognized that because DVTS software was not installed in PC11, PC11 cannot play DVTS video. Consequently, the USB camera connected to PC1 with the JMF-send Ag is selected. In this case, we confirmed that our scheme could deeply consider the multiple contexts, and our scheme could satisfy the user requirement for high smoothness of the video.

6.2 Performance Evaluation

We performed an additional experiment to measure the switching time during video service migration for performance evaluation of our system. We used the user terminal with two kinds of access networks: IEEE 802.11g (54 Mbps) and PHS (128 kbps) in this experiment. We measured the switching time during the video service migration in the cases of IEEE 802.11g and PHS, respectively.

We measured two cases of switching time.
Case-1: The video service migrated from the user terminal (PC12) to the other PC based on the user request.
Case-2: The video service migrated from the average PC except the user terminal to the other PC (except the user terminal).

When the hand-held PC used IEEE 802.11g, both Case-1 and Case-2 were able to switch within 3.0 s, on average. It is in an acceptable range for practical use. When the hand-held PC used PHS, Case-1 took more than 7.0 s in some cases, but Case-2 switched within 3.0 s on average. This reason of the time delay in Case-1 with PHS is a latency of inter-agent message exchange for video service migration between agents in the hand-held PC and the other PC, during which time the hand-held PC is receiving the video streaming. On the other hand, the switching time in Case-2 with PHS was almost the same as in those cases using IEEE 802.11g. This result shows that the individual agent could decide its own action by considering the situations of the other agents.

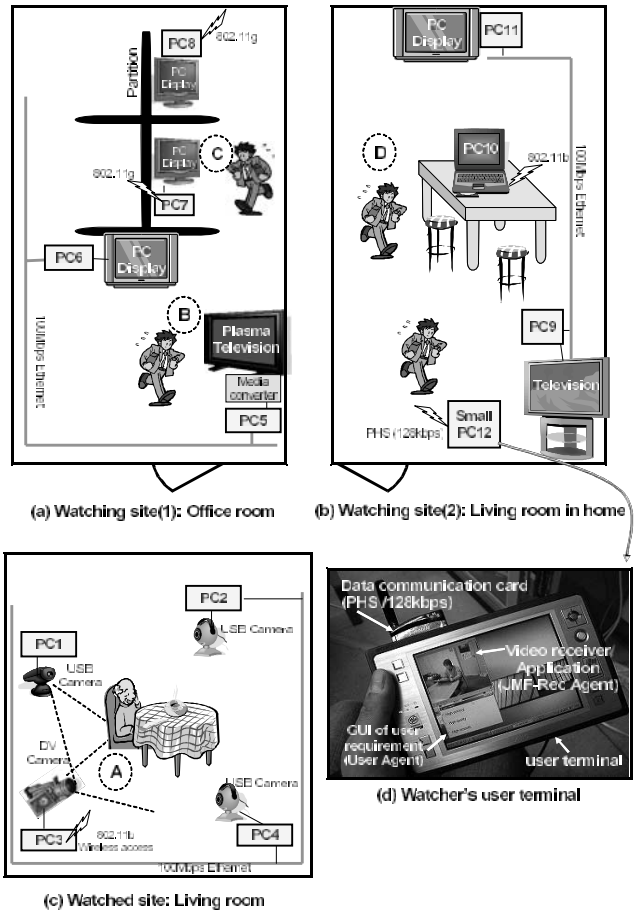


Figure 4: Experimental room settings

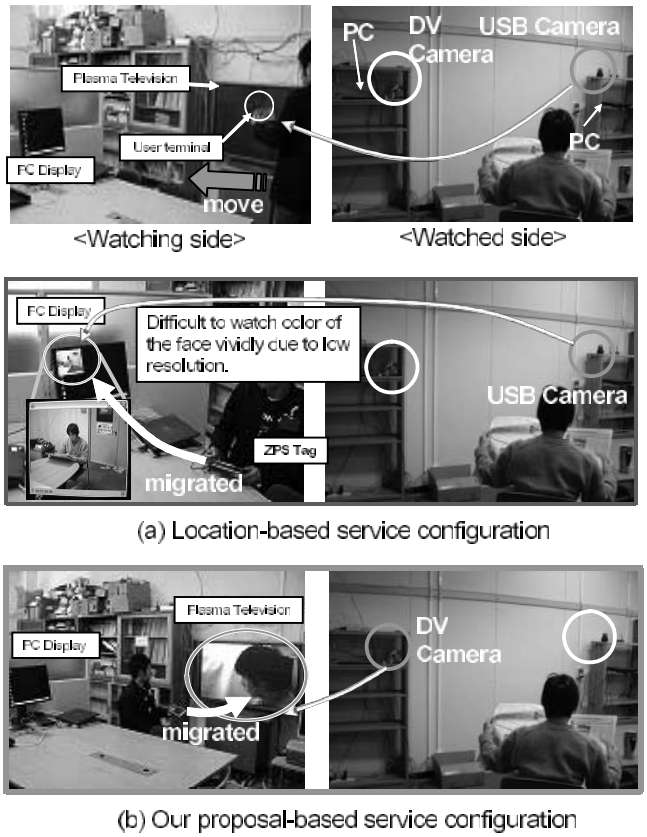


Figure 5: Service configuration in case of high quality requirement at point "B" in Figure 4(a).

6.3 Privacy protection function by controlling quality of service

We think privacy concerns are important aspect in healthcare support system. We are now trying to give privacy protection function to our system. Figure 7 shows the privacy protection function using JMF by controlling the level of quality. This function adjust the parameters related the video quality of JMF such as frame rate, bit rate, etc., in accordance with the video format (Motion-JPEG and H.263). In fact, JMF-send agent and JMF-rec agent cooperate to adjust the parameters and the format depending on the situation of network resource. Figure 7(a) shows the case of Motion-JPEG; Figure 7(b) shows the case of h.263. We can see from Figure 7(a) and Figure 7(b), the quality of the video is too low to see the person's face clearly, but we can only judge the person's movement. We consider this function can useful as one of the method to protect the privacy easily.

6.4 Discussion

We discuss the effectiveness of our proposal through the experiments described in the previous sections.

- **Feasibility:** We evaluated our proposal-based service configuration scheme. Our scheme could effectively configure service that matches person's requirement, coping with not only the user location information, but also the device status around the users in the ubiquitous computing environment. In this real-time multimedia supervisory function based on our healthcare system, heterogeneous entities like display devices, capture devices, PCs, networks, different kinds of sensors, software components, etc., are efficiently integrated.

- **Effect of Framework:** Because of the introduction of layered software architecture with the entities and agents, the integration of many entities was successful. This architecture is employed to accelerate the reuse of the framework and the middleware by other types of applications. The modularity, the autonomy, and the loose coupling characteristics of the agents would meet to the construction of various healthcare support systems. It can adapt to diversity of types of entities and scalability of system size. The system development and extension will be easily accomplished by using this architecture.

7 CONCLUSION

We presented a concept of user-oriented advanced healthcare support system based on multi-agent system in ubiquitous computing environment. The system provides useful information regarding health condition effectively and in user-

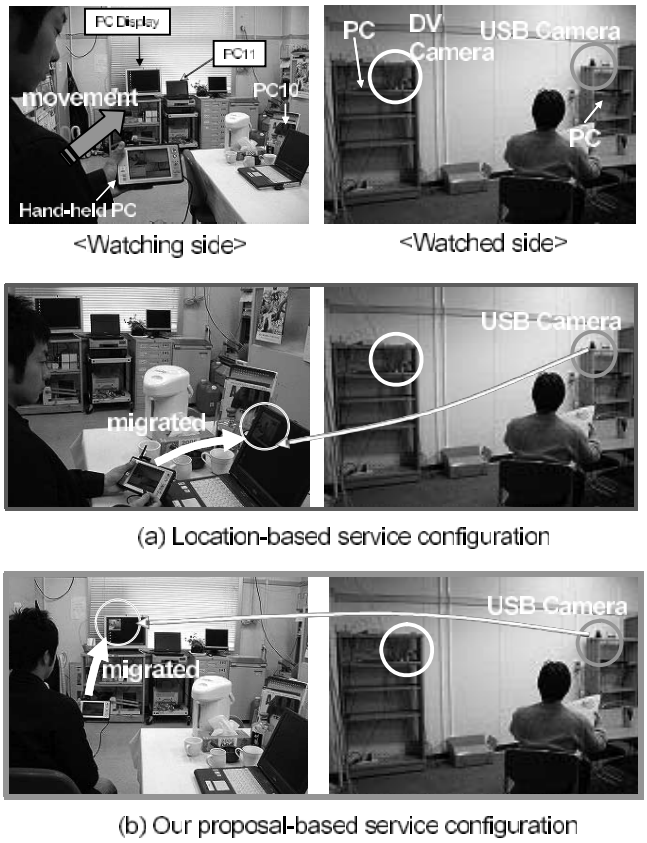


Figure 6: Service configuration in case of high smoothness requirement at point "D" in Figure 4(b).

oriented manner by utilizing knowledge about healthcare and various kinds of information obtained from real space. We also designed and implemented an initial prototype system.

As future work, we would like to advance detail modeling and design to adapt to a variety of the supervisory system such as the healthcare support system and multimedia watching over system for elderly people, and we plan to extend current implementation so that it can be enable various evaluation.

ACKNOWLEDGEMENT

This work was partially supported by Sendai Intelligent Knowledge Cluster and the Ministry of Education, Culture, Sports, Science and Technology, Grants-in-Aid for Scientific Research, 19200005.

REFERENCES

[1] http://www.mhlw.go.jp/bunya/kenkou/seikatsu/index.html.
[2] http://www.health.ne.jp/.
[3] http://www.ch-kentatsu.jp/.
[4] E. Hyvonen, K. Viljanen, and O. Suominen, HealthFinland - Finnish Health Information on the Semantic Web, ISWC/ASWC 2007, LNCS 4825, pp. 778–791(2007).

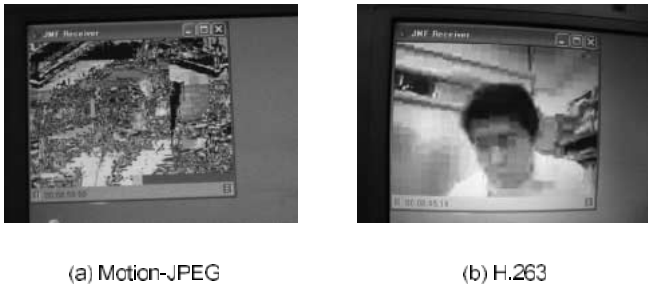


Figure 7: Privacy protection function of video streaming in case of Motion-JPEG (a) and H.263 (b)

[5] http://air.silpheed.jp/users/sfais/0.php.
[6] S. Izumi, Y. Kato, K. Takahashi, T. Suganuma, and N. Shiratori, Proposal of an Ontology-based Health Support System and Its Evaluation IPSJ Journal, Vol. 49, No. 2, pp. 822–837(2008).
[7] D. Noheji, G. Itabashi, T. Yoshida, and A. Togashi, Human Resource Management System introducing Healthcare of Employee, ISWC2007+ASWC2007 Poster+Demo Proceedings, pp. 79–80(2007).
[8] U. Varshney, Pervasive Healthcare and Wireless Health Monitoring, ACM/Baltzer Journal of Mobile Networks and Applications, Vol. 12, No. 2-3, pp. 113–127(2007).
[9] B. Brown, M. Chetty, A. Grimes, and E. Harmon, Reflecting on Health: A System for students to monitor diet and exercise, Proc. of Conference on Human Factors in Computing Systems (CHI'06), pp. 1807–1812(2006).
[10] R. Gockley, M. Marotta, C. Rogoff, and A. Tang, AVIVA: A Health and Fitness Monitor for Young Women, Proc. of Conference on Human Factors in Computing Systems (CHI'06), pp. 1819–1824(2006).
[11] D. Kuriyama, S. Izumi, G. Itabashi, S. Kimura, Y. Ebihara, K. Takahashi, and Y. Kato, Design and Implementation of a Health Management Support System Using Ontology, Proc. of the International Conference on Engineering, Applied Sciences, and Technology (ICEAST 2007), pp. 746–749(2007).
[12] D. Cruz, and E. Barros, Vital Signs Remote Management System for PDAs, Proc. of the 2005 8th Euromicro Conference on Digital System Design, pp. 170–173(2005).
[13] R-M. Dröes et al., Healthcare Systems and Other Applications, IEEE Pervasive Computing, Vol. 6, No. 1, pp. 59–63(2007).
[14] K. Chang, M.Y. Chen, and J. Canny, Tracking Free-Weight Exercises, UbiComp 2007, LNCS 4717, pp. 19–37(2007).
[15] K. Ouchi, T. Suzuki, and M. Doi, LifeMinder: A Wearable Healthcare Support System with Timely Instruction Based on the User's Context, IEICE Transaction on Information and Systems, Vol. E87-D, No. 6 pp. 1361–1369(2004).
[16] E.J. Ko, H.J. Lee, and J.W. Lee, Ontology-Based Context Modeling and Reasoning for U-HealthCare, IEICE

Transactions on Information and Systems, Vol. E90-D, No. 8, pp. 1262–1270(2007).
[17] N. Ali-Hasan, D. Gavales, A. Peterson, and M. Raw, Fitster: social fitness information visualizer, Proc. of Conference on Human Factors in Computing Systems (CHI'06), pp. 1795–1800(2006).
[18] M. Kudlacz, R. Tan, J. Prindiville, and M. Peters, RoutePlanner, Proc. of Conference on Human Factors in Computing Systems (CHI'06), pp. 1849–1854(2006).
[19] Symbiotic Computing Home Page, http://symbiotic.agent-town.com/
[20] T. Suganuma, T. Uchiya, S. Konno, G. Kitagata, H. Hara, S. Fujita, T. Kinoshita, K. Sugawara, and N. Shiratori, Bridging the E-Gaps: Towards Post-Ubiquitous Computing, Proc. of the 1st International Symposium on Frontiers in Networking with Applications (FINA 2006), pp. 480–484(2006).
[21] T. Suganuma, K. Sugawara, and N. Shiratori, Symbiotic Computing: Concept, Architecture and Its Applications, Proc. of the 4th International Conference on Ubiquitous Intelligence andComputing (UIC-07), pp. 1034–1045(2007).
[22] H. Takahashi, Y. Tokairin, T. Suganuma, and N. Shiratori, Design and Implementation of An Agent-based middleware for Context-aware Ubiquitous Services, Frontiers in Artificial Intelligence and Applications, New Trends in Software Methodologies, Tools and Techniques, Vol. 129, pp. 330–350(2005).
[23] H. Takahashi, T. Suganuma, and N. Shiratori, AMUSE: An Agent-based Middleware for Context-aware Ubiquitous Services, Proc. of the International Conference on Parallel and Distributed Systems (ICPADS2005), pp. 743–749(2005).
[24] S. Fujita, H. Hara, K. Sugawara, T. Kinoshita, and N. Shiratori, Agent-Based Design Model of Adaptive Distributed Systems, Applied Intelligence, Vol. 9, pp. 57–70(1998).
[25] K. Sugawara, H. Hara, T. Kinoshita, and T. Uchiya, Flexible Distributed Agent System programmed by a Rule-based Language, Proc. of the Sixth IASTED International Conference of Artificial Intelligence and Soft Computing, pp. 7–12(2002).
[26] T. Uchiya, T. Maemura, L. Xiaolu, and T. Kinoshita, Design and Implementation of Interactive Design Environment of Agent System, Proc. of the 20th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE 2007), pp. 1088–1097(2007).
[27] Zone Positioning System, http://www.furukawakk.jp/products/
[28] Local Positioning System (LPS), http://jp.fujitsu.com/group/fst/services/ubiquitous/rfid/
[29] A. Ogawa, K. Kobayashi, K. Sugiura, O. Nakamura, and J. Murai, Design and Implementation of DV based video over RTP, Packet Video Workshop 2000(2000).

Research about Method of Presuming Incident in Information Security

Masahiko Ishino*, Naokazu Yamaki**, Tsukasa Kudou*** and Kenji Saotome****

* Mitsubishi Electric Information Technology Corp. Japan
** Faculty of Engineering Shizuoka University Japan
*** Mitsubishi Electric Information Systems Corp.
**** Hosei Business School of Innovation Management Japan

Abstract - Many information securities measures are being needed by development of the internet society. We propose a method to presume the incident about the information securities in each section of the enterprise. We presented research about the algorithm of the recommendation system. We presume the incident which can occur now to IT (Information Technology) assets and characteristic of each section using the method. We were changed attributes of the warning log data which occurred on each section of the enterprise and presume the similar incident which doesn't occur yet. We can give priority to the information securities measure to the incident which presumed the possibility which occurs to be high by applying log data to this method.

Keywords: Information Security, Recommendation System, Analytic Hierarchy Process (AHP), Conjoint Analysis, Collaborative Filtering, Data-mining,

1 INTODUCTION

The new big wave of Web2.0 came to society on now, by the internet spread all over the world and information technology developed. We got possible to enjoy cheap, convenient much service that used the internet in society. However, on the other hand, the information security is threatened every day. For example, we are exposed to uneasiness that there is unjust access to the computing system, the virus spreads by the internet, and important information is stolen, or leak out the information security from survey data in Figure 1-3. The source of Fig 1-3 are the investigation report about the information security incident in 2003 of Japanese Network Security Association

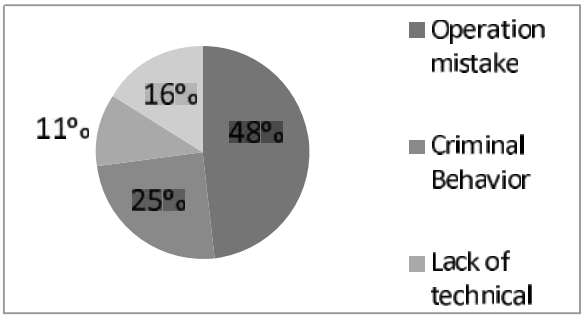


Figure 1: Main cause of information leak

In the enterprise, the incident about the information security occurs frequently year by year and makes efforts in those measures. According to the investigation report about the information security incident in 2003 of Japanese Network Security Association (JNSA), the main cause of the information leak is operation error (48%) and crime operation (25%). The main exit of the information leak is Web(20%), E-mail (17%), USB memory (16%),paper (14%).

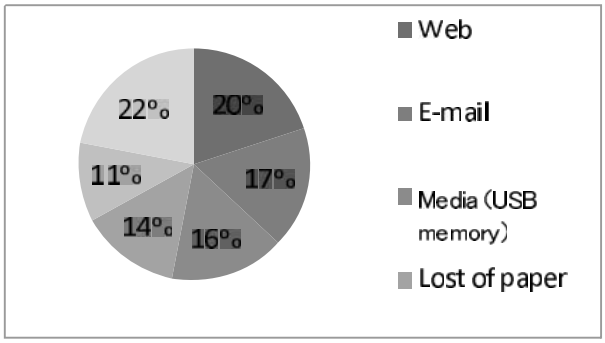


Figure 2: Main exit of information leak

In addition, uneasy item about the information security in the enterprise is incompleteness of the anti-security measure (54%). The uneasy item by the information leak is unjust carrying out (35%), leak from the trust of the employees (25%). When the information leaked out, there is the damage on the brand of the enterprise (74%), business suspension of the license (9%), and the suit risk (7%). The enterprise (27%) expects education about the information security for the employee.

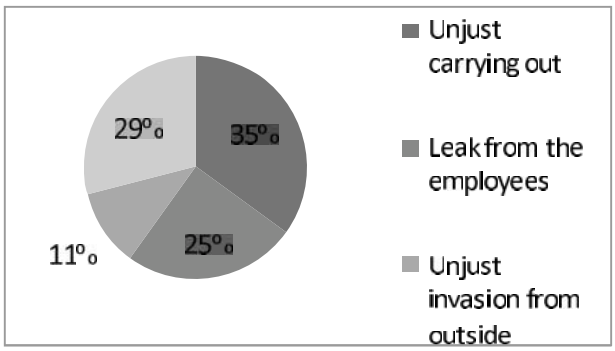


Figure 3: Uneasy item about the information security

At first the purpose of the presuming incidents in the information security is prediction of the incident that can occur for the attribute of IT assets of each section as we show it in Figure 4.

We apply the method of recommendation system which fitted the taste of the customer to the presuming incidents. The system converts the log data of the incident that occurred into many sections to the attribute. By the method we presumed the similar incident that does not yet occur in the specific section. If it is very likely that it occurs from the presumed result, the anti-information security measure for the presuming incident is possible.

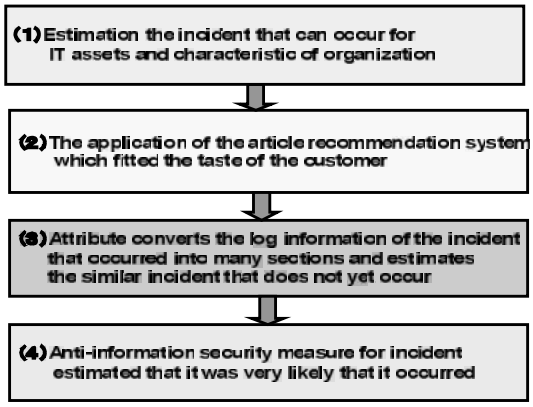


Figure 4: Purpose of the presuming incidents

In this research, we applied the technique that published the results of our research about recommendation method by the mathematics model until last year to the estimate method of the incident in the information security [1], [2]. We built mathematics model and utilized collaborative filtering and conjoint analysis and data-mining, the division by Analytic Hierarchy Process (AHP) and suggested the choice standard of the recommendation candidate and estimate and article recommendation method of the article taste evaluation value so far [3], [4]. In this presentation of the results of the research, we suggest the method of the presuming incident in the information security.

2 COMPARISON WITH SIMILAR RESEARCH

There is "Applications of Mining to Information Security" by Kenji Yamanishi with the representative similar research [5]. The purpose is building of security intelligence, and the research is detected abnormally in real time from access record. He analyzed an incident using standard statistics technique in the research. Because the method uses many data, there is much computational complexity. It is easy to use our method by the simple Computational complexity by the attribute matrix in comparison with the precedent similar research. Our method can narrow down the anti-information security measure of the presumed result easily.

The incident estimate system carries out the measures of the incident from log analysis to show it in Figure 5 and repeats cycle to feed back from the result to the next estimate.

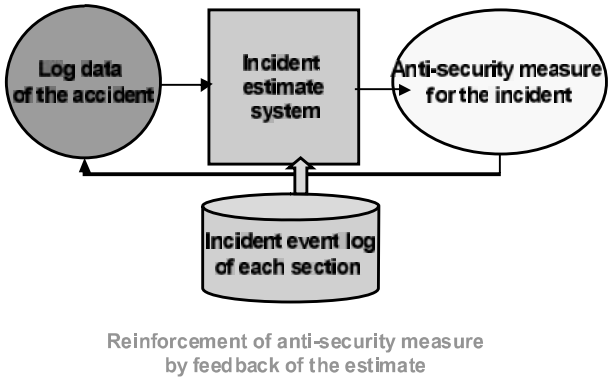


Figure 5: Anti-security measure by incident estimate

3 PRESUMED METHOD OF INCIDENT

This method defines the mathematics model of "Attribute Relations Matrix" that express the size of the characteristic at weight (It is called " attribute" as follows) of the section as the characteristic of various kinds of incidents and will set the estimate value of assumed incident in future in section of certain characteristic. This method chooses the attribute of incident from much log data to show it in Figure 6. Furthermore, the anti-information security measure is narrowed down.

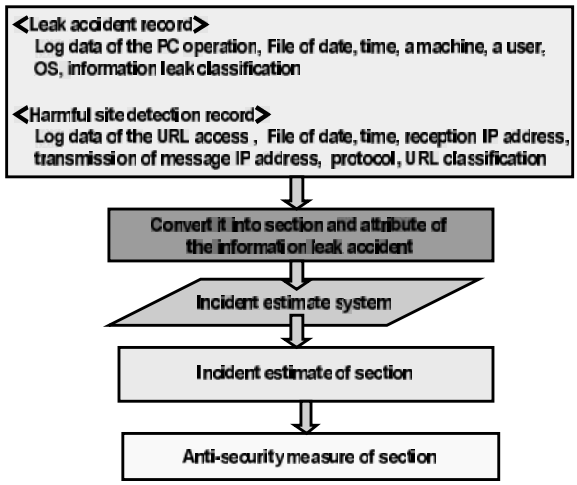


Figure 6: Anti-security measure by information system

3.1 Kind of incident

We have various factors and measures for the incident about the information security. Those classifies particularly main factors is shown in Table1. The enterprise performs various plan and introduction of many measures systems for these menaces in the enterprise. We estimate the high incident of the priority from the log data of the other section to carry out the measures that matched the attribute of the section of the enterprise.

We estimate incident that can occur from the attribute of the warning log data of each sections. For example, there are outside unjust invasion, intentional outside unjust invasion inside information leak system destruction data manipulation unjust copying of the classification artificial of main incident factor in Table 1.

Table 1: Classification of main incident factor

Classification	Artificial factor	Non-artificial factor
Intention	• Outside unjust invasion • Internal information leak • System destruction • Data manipulation • Unjust copying, etc.	.....
Fault	• Mistakes of operation • Data loss • Data non-elimination • System malfunction	• Natural disaster • Infrastructure disorder except system

3.2 Kind of log data

We gather warning log from various log data such as Table 2 in the information system of the enterprise routinely and investigate cause at the time of accident of the information security.

We usually gather various log data by the computing system automatically. We plan the measures of artificial incident for intention in these. However, it is generally difficult for the tendency to security risk to get it because quantity of data is enormous.

Table 2: Kind of log data

Object system	Kind of the log
Operating system	Event log
Database system	Access log
Network system	Access log
E-mail	Transmission / reception log
Web system	Access log
Server system	Access log
Terminal system	Operational log
Internet search	Access log

3.3 Attribute value vector

We use conjoint analysis for the incident attribute and the section attribute of Attribute Relation Matrix and perform the weight charge account. And we use the weight and apply it to the estimate of the incident to the section. We set the classification (unjust invasion, information leak, unjust copying, operation mistake, others) of the incident and the attribute of time range and converted all incidents from warning log according to the section into the attribute value vector of the incident. I perform it in conjoint analysis about the tendency of the incident attribute every section by the outbreak number of the warning log of the identification period and do a weight charge account for the incident attribute. Each section sets duties, IT assets, an equal ∅ attribute and converts log data according to the incident into the attribute value vector of the section. WE consider that a section with much outbreak frequency of the warning log is very likely to be the future, the incident outbreak and, by conjoint analysis, perform the weight charge account of the section attribute. For example, attribute vector of incident is shown in equation (1).

$$x_j, j = 1 \sim m \in B^m \quad (1)$$

And, attribute vector of section is shown in equation (2).

$$y_i, i = 1 \sim n \in B^n \quad (2)$$

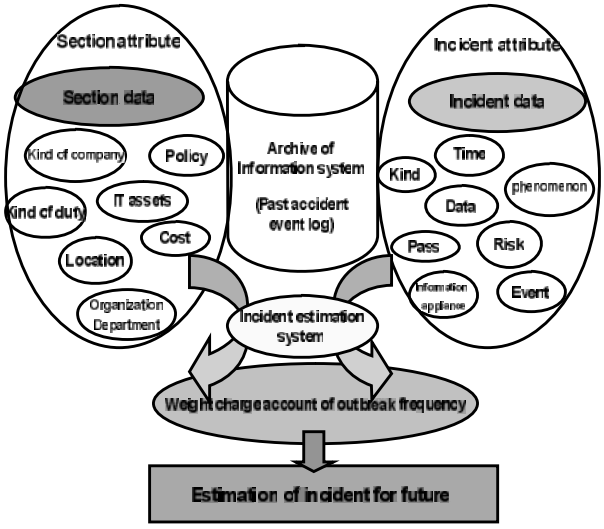


Figure 7: Example of Section and Incident Attribute

From attribute group such as Figure 7, we exemplify Incident Attribute Matrix consisting of the incident attribute vector in Table 3. We insert "1" in the column to fall under every incident by classification and the outbreak time of the attribute and generate Incident Attribute Matrix. Section Attribute Matrix consisting of the section attribute vectors inserts "1" in the column falling under duties and IT assets every section and generates Section Attribute Matrix likewise such as Table 4.

Table 3: Incident Attribute Matrix

Attribute	Kind of incident					Time range		
Kind of Incident	Unjust access	Unjust copy	Copy on information	Non Code	Mistake in operation	Day time	Night	Holiday
a_1	1	0	0	0	0	0	1	0
a_2	0	0	1	0	0	1	0	0
a_3	0	0	0	0	1	0	1	0
⋮								
a_k	0	1	0	0	0	1	0	0

In other words the incident to fall under each attribute is classified in some patterns.

Table 4: Section Attribute Matrix

Attribute	Kind of job				IT asset				
Section Name	General affairs	Sales	Production	Research and development	Personal information	Design information	Sale information	Intellectual property	...
d_1	1	0	0	0	1	0	0	0	
d_2	0	0	1	0	0	1	0	0	
d_3	0	1	0	0	0	0	1	0	
⋮									
d_l	0	0	0	1	0	0	0	1	

3.4 Setting of weight for attribute value vector

We express the part-worth matrix of the incident attribute for the object section which demanded incident attribute vector and each of the weight charge account of the outbreak order of the section attribute in the identification incident which express weight charge account of the outbreak order of the incident attribute in the specific section by conjoint analysis in equation (3).

$$U_i \in R^{n \times m} \tag{3}$$

The weight matrix of the incident for the section becomes equation (4).

$$U_i = (u_{j1}, u_{j2}, \cdots, u_{jm}) \tag{4}$$

The presumed evaluation value of the incident for the section is expressed in equation (5).

$$E_{a_j} = U_i^T x_j \tag{5}$$

Using the weight matrix in search of the outbreak evaluation value of the incident, do a high incident of the presumed evaluation value with the choice candidate of the estimate incident from an incident set.

The weight matrix of the section attribute is shown by equation (6) for the identification incident.

$$V_j \in R^{m \times n} \tag{6}$$

The weight matrix of the section for incident a_j is shown by equation (7).

$$V_j = (v_{i1}, v_{i2}, \cdots, v_{in}) \quad (7)$$

The presumed evaluation value of the section d_i for the incident a_j is expressed by equation (8).

$$E_{d_i} = V_j^T y_i \quad (8)$$

Attribute Relation Matrix $W^{m \times n}$ that weight setting to relations from two attributes of incident and section, defined by next equation (9).

$$W = \begin{bmatrix} w_{11}, w_{12}, w_{13}, \cdots, w_{1n} \\ w_{21}, w_{22}, w_{23} \cdots, w_{2n} \\ \vdots \qquad\qquad \vdots \\ w_{m1}, w_{m2}, w_{m3}, \cdots, w_{mn} \end{bmatrix} \begin{matrix} \text{Incident Attribute} \\ \\ \text{Section} \\ \text{Attribute} \end{matrix} \quad (9)$$

In addition, Attribute Relation Matrix $W^{m \times n}$ is led according to the Figure 8 of the method to generate $W^{m \times n}$ concerned.

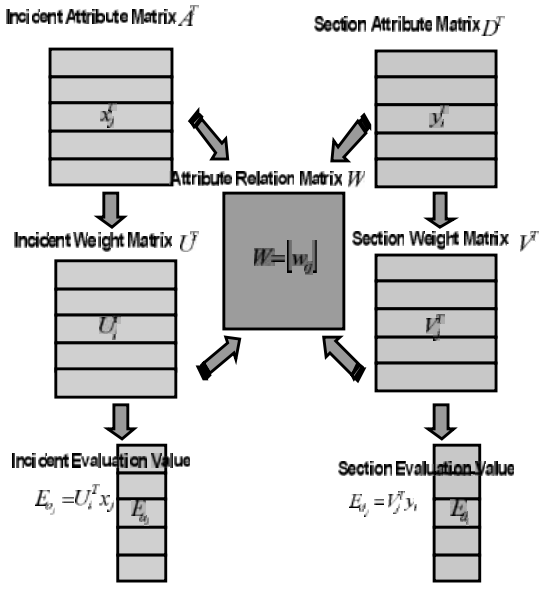


Figure 8: Constitution of Attribute Relation Matrix

This Attribute Relation Matrix $W^{m \times n}$ and weight matrix of incident attribute is demanded from the specific section attribute vector, and weight matrix of the section attribute is demanded by specific incident attribute. The incident weight matrix of section becomes to equation (10).

$$U_i = W \quad y_i \quad (10)$$

Incident presumed evaluation value E_{d_i} is provided fromnext equation (11).

$$E_{d_i} = U_i^T x_j \quad (11)$$

The handling of technique procedure becomes it according to the following Figure 9. We used collaborative filtering and conjoint analysis by a process of log analysis.

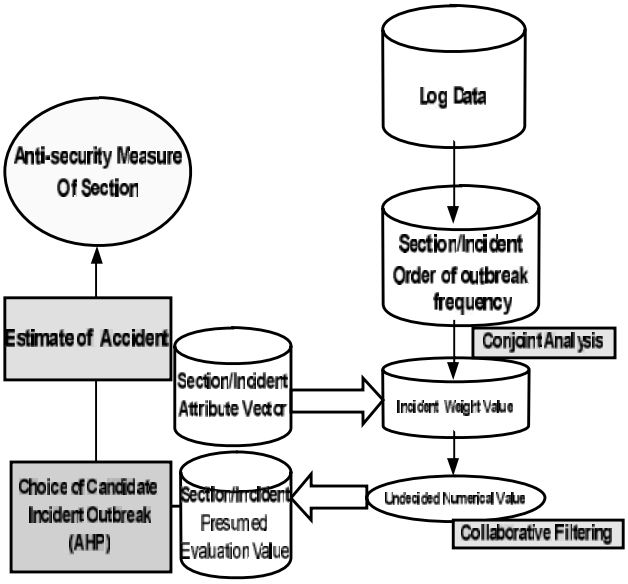


Figure 9: Data conversion flow

Anti-security measure is chosen among this result of log analysis by AHP method. We repeat a PDCA cycle of measures and the practice and raise presumed precision.

Here, exemplify the weight matrix of the incident attribute and presumed evaluation value in Table 4. Presumed evaluation value takes first priority from a high incident and examines measures.

Table 4: Presumed evaluation value of incident attribute

Kind of Incident	Unjust access	Unjust copy	Carry on Information	Non Code	Mistake in operation	Day time	Night	Holiday	Presumed evaluation value
a_{r1}	6	0	0	0	0	0	4	0	10
a_{r2}	0	0	5	0	0	3	0	0	8
a_{r3}	0	0	0	0	4	0	3	0	7
⋮									
a_{rk}	0	2	0	0	0	1	0	0	3

The weight matrix of the section according to the incident becomes equation (12) likewise.

$$V_j = W^T x_j \qquad (12)$$

3.5 Necessity of AHP

More than person's choice as well as detection of the automatic incident attribute were taken in. This reason is necessity of most suitable AHP (Analytic Hierarchy Process).

We enumerate presumed incidents from the order of the presumed incident that we demanded by this technique, and it is possible for narrowing by the evaluation standard of ISMS by AHP to resemble Figure 10 in a flow.

4 THE APPLICATION OF LOG DATA

We applied this technique from the log data which gathered from information system in the enterprise. The log data of the object extracted warning log data except normal log data in the system use and found incident presumed weight value. We generated an attribute-related matrix from Section Attribute Matrix and Incident Attribute Matrix under the following condition.

4.1 Selection of Attribute

We chose the following attributes with the method.

(1)Duties classification : General affairs, business, production, research and development

(2)IT assets: Personal information, design information, sale information, intellectual property

(3)Data : Warning log data

(4)Period : One month,

(5)Incident classification : Unjust access, Unjust copying, Information carrying out, Non-Coding, Operation mistake

(6)Time: Day time , Night, Holiday

4.2 Procedure of Incident Estimate Method

Preparations and the processing procedure of data are as follows.

(1)The preparations for data

- Choice of the object section
- Extraction of warning log

(2)Data processing

- Choice of attribute
- IT assets of section
- Outbreak frequency
- Presumed weight
- Attribute Relation Matrix

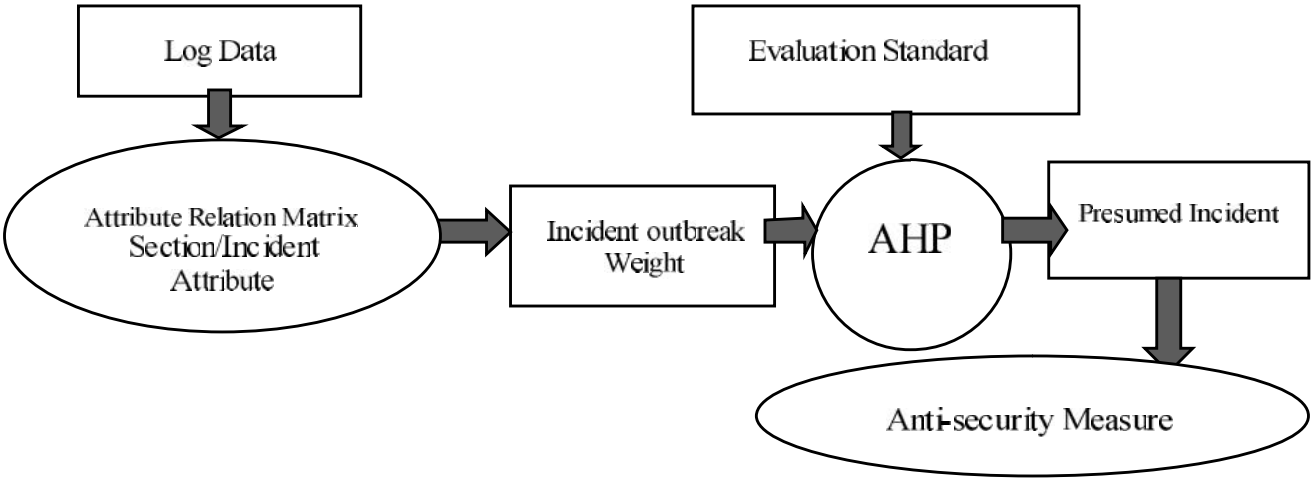


Figure 10: Choice flow of incident estimate

The weight value between are generated such Attribute Relation Matrix as Table 5.

Table 5: Attribute Relation Matrix

	Attribute	Kind or incident					Time Range		
		Unjust access	Unjust copy	Carry of information	Code of non-processing	Mistake in operation	Day time	Night	Holiday
Section	General affairs	4	3	5	2	1	2	1	1
	Sales	1	2	1	1	4	3	1	1
	Production	2	1	2	1	1	2	1	3
	Research and development	3	4	1	2	1	2	3	1
IT Asset	Personal information	2	1	3	1	5	2	4	1
	Design information	4	1	4	3	1	1	1	1
	Sale information	1	4	2	2	3	1	2	1
	Intellectual property	2	2	3	1	1	2	4	1

The evaluation estimate p_i the target of incident in the section d_i. It is defined as equation (13).

$$p_i = W^T y_i \qquad (13)$$

We defined the attribute vector of the section d_i of general affairs, and personal information to equation (14),

$$y_i = (1,0,0,0,1,0,0,0) \qquad (14)$$

The presumed evaluation value vector of the incident in the section becomes equation (15) in Table 6.

$$p_i = W^T y_i = (6,4,8,3,6,4,5,2)^T \qquad (15)$$

From equation (15), the outbreak estimate weight value of the incident carry of information is high in night time range.

The evaluation estimate vector of the section for the incident a_j is demanded as equation (16).

$$q_j = W^T x_j \qquad (16)$$

When we sentenced the attribute vector of the incident of unjust copy to equation (17) in night time of the incident,

$$x_j = (0,1,0,0,0,1,0,0) \qquad (17)$$

Table 6: Evaluation value of Incident Attribute

Section Attribute: (1, 0, 0, 0, 1, 0, 0, 0) ×

Attribute	Unjust access	Unjust copy	Carry of information	Code of non-processing	Mistake in operation	Day time	Night	Holiday
General affairs	4	3	5	2	1	2	1	1
Sales	1	2	1	1	4	3	1	1
Production	2	1	2	1	1	2	1	3
Research and development	3	4	1	2	1	2	3	1
Personal information	2	1	3	1	5	2	4	1
Design information	4	1	4	3	1	1	1	1
Sale information	1	4	2	2	3	1	2	1
Intellectual property	2	2	3	1	1	2	4	1

= [6,4,8,3,6,4,5,2]

Weight of incident attribute Information carrying out, Night range

The presumed evaluation value vector of the section for the incident is shown as equation (18) in Table 7.

$$q_j = W^T x_j = (5,5,3,6,3,2,5,6)^T \qquad (18)$$

Table7. Evaluation value of Section Attribute

Attribute	Unjust access	Unjust copy	Carry of information	Code of non-processing	Mistake in operation	Day time	Night	Holiday
General affairs	4	3	5	2	1	2	1	1
Sales	1	2	1	1	4	3	1	1
Production	2	1	2	1	1	2	1	3
Research and development	3	4	1	2	1	2	3	1
Personal information	2	1	3	1	5	2	4	1
Design information	4	1	4	3	1	1	1	1
Sale information	1	4	2	2	3	1	2	1
Intellectual property	2	2	3	1	1	2	4	1

× Incident Attribute: (0, 1, 0, 0, 0, 1, 0, 0) = [5, 5, 3, 6, 3, 2, 5, 6]

Section attribute: Research and development (6); Intellectual Property (6)

From equation (18), the outbreak estimate weight value of Section Attribute of R&D, the intellectual property is high. We assume it candidate choice of the presumed incident from the section attribute that such the weight value is high.

5 CONSIDERATION

In this research, we applied log data for one part from information system in the enterprise and considered the presumed method of the incident about the information security.

(1) With the attribute-related matrix, we suggested it as a method of the one which estimated the estimate of the incident for the section with the certain attribute and the section for the incident of the certain attribute.
(2) Because it is estimate, we compare it with analysis after the thing of log data, and this technique can estimate other similar incidents by the incident attribute from the incident attribute and the section attribute.
(3) It repeats itself and, as well as the estimate of the incident that applied the log data of the identification period, applies log data, and information security measures of the type are necessary for PDCA cycle estimating an incident.

6 CONCLUSION

We show the conclusion of the research next. There is the effectiveness of new method in our research. The method can offer the next merit.

(1) Realization to visible of the incident
(2) Minimization of the computational complexity
(3) Extraction of the attribute of the incident of section
(4) Novelty of the matrix method
(5) Support to the anti-security measure

The next Figure11 shows the unification log analysis system. We implement the algorithm that we used suggestion technique for practical use to general log system of administration. This system can support measures for various incidents from the unification log analysis database.

The interval of the log data upload and the frequency of the analysis change by this system are plan to the optimization. The human interface by visual expression of the data analysis is important.

We will perform the following studies about the incident estimate method in next our research.

(1) The section attributes (section, IT assets) and the incident attributes (incident classification, time range) from log data, we examine the extraction method of other attributes.
(2) We push forward the research about the uptake method of chronological order data corresponding to the change of the log of the specific section.

We investigate data-mining of precedent research about application to information security and will refer to it in future. We want to perform the further research about practical method about the estimate of the incident attribute.

REFERENCES

[1] Resnick P.,and Varian, H. R., "Recommender Systems," Comm. of the ACM, Vol.40, No.3, pp.56-58, 1997.
[2] P. Resnick, "GroupLens, An Open Architecture for Collaborative Filtering of Netnews." Proceeding of the Conference on Computer Supported Cooperaive Work (CSCW'94),1994,pp.175-186.
[3]T.L. Saaty, "The Analytic Hierarchy Process", RWS Publications, 1990.
[4]Michael J.A.Berry and Gordon S .Linoff, "Data Mining Techniques:For Marketing,Sales,and Customer Relationship Management.",2nd Edition WilePublishing,Inc. ,Indianapolis,Indiana
[5]Kenji Yamanishi, Applications of Mining to Information Security, "Information Security and Artificial Intelligence", Journal of Japanese Society for Artificial Intelligence,Vol.21,No.5,pp.570-576

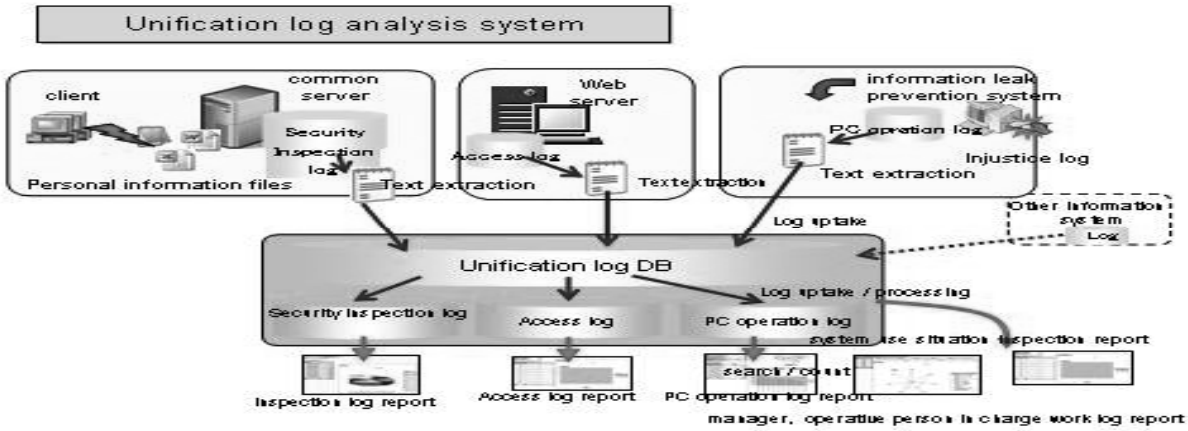


Figure 11: Unification log analysis system