

Ubiquitous Data Management by using Word-of-mouth Information

Yasuyuki BEPPU and Toshio TONOUCHI

Internets Systems Research Labs. NEC Corporation,

1753 Simonumabe, Nakahara-Ku, Kawasaki,
Kanagawa, 211-8666 Japan

E-mail:{y-beppu@ak, tonouchi@cw}.jp.nec.com

Abstract

These days, people can obtain variety of information on any subject by accessing the Internet. Because the amount of available information existing on the Internet is enormous, it is difficult to find relevant information. However, information from friends with the same interests should be valuable and trustworthy.

Web sites accessed by community members with the same interests can include a greater proportion of relevant contents for other members than ranked web sites in which people obtain information using a keyword searching engine. We developed a management software, called a word-of-mouth information management system that, enables community members to get the URLs from other community members and to supply them to other members without taking any additional actions.

Nine people took part in experimental evaluation of the system and accessed 167 contents on the Internet by using our system. Our results verified that the tool is valuable for finding relevant information over the Internet.

Keywords

Ubiquitous data management, word-of-mouth, P2P, and XML

Introduction

- In the ubiquitous age, information available on the Internet is huge and to find **actually useful** information is difficult.
- **Word-of-mouth** information from friends having same interests is **valuable and trustworthy**.
- We develop and evaluate a management system which enables to **share URLs** of web sites among friends.
- The most important issue is that distributed information includes **valuable** one for receivers.

1. Introduction

The ubiquitous age, in which people will have access to the Internet anytime-anywhere, is coming soon. People can easily record information on homepages or blog sites. The amount of information available over the Internet is huge and increases everyday. There is also a lot of erroneous information available. Therefore it is difficult to find relevant and reliable information quickly.

These days, word-of-mouth information from friend with the same interests is considered to be valuable and trustworthy. In fact, marketing results show that the primary motivation for buying is word-of-mouth information for almost all products [1].

Distributing information this way is effective in the marketing world. Web sites accessed by community members that share the same interests can include a greater proportion of relevant contents for a member of than ranked web sites in which people obtain information using a keyword searching engine. The reason is that people with the same interests have likely already checked the sites, filtered the contents, and have judged the sites to have a value to look at. Therefore, by exchanging web site URLs of web, people who share the same interests, can obtain looked valuable and trustworthy information.

We developed a management software, called the word-of-mouth information management system that, enables people to transmit information from their PC to the PC of another community member using word-of-mouth method. The first and most important issue is to develop a management system to distribute information relevant to the receivers. There are two other issues to use the management system effectively. One is that people need to exert little effort to share word-of-mouth information. The other issue is that the ratio of relevant information distributed, which is defined as the ratio of the number of relevant information units for receivers to that of all the information units distributed by our management system should be high, because it enables users to obtain relevant information units in short time. We confirmed the effectiveness of our system in an evaluation experiment.

Background: Word-of-mouth and P2P

- P2P architecture that exchange information PCs directly is paid attention. Napster [Lechner, 2001], Gnutella [Ripeanu, 2001], Skype [Salman, 2004]
- Word-of-mouth and P2P have common characteristics in information transmission method.
- To develop a system transmitting word-of-mouth information, therefore, P2P architecture is more suitable than client-server architecture.

2. Background and Approach

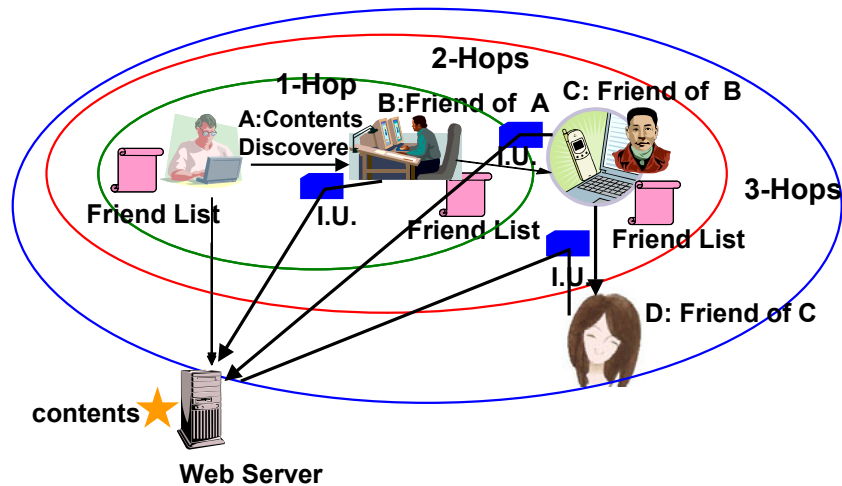
2.1 Background

In the world of information systems, in addition to typical client-server architecture that manages using central servers, we focus on peer to peer (P2P) [2] architecture that does not depend on specific servers and still enables PCs to communicate each other directly. Some P2P type applications have been developed recently. The major application software includes Napster [3], which enables the exchange of MP3 format files by supplying directory service at central server, Gnutella [4], which enables general file exchanges between peers in a P2P network without depending on central servers, and Skype [5] that supplies voice over IP networks at a low cost.

Each PC, which distributes information using word-of-mouth over information systems, distributes an information units to other PCs and they distribute information units in succession. In general, PCs that cannot identify the PC that got the information first. This system architecture is suitable because it can transmit information to PCs over a large area, which includes the friend of friend's PC from the originating PC. In general, by tracing the steps from one friend to another friend, the amount of information transmitted to the next person usually decreases. But in this case, the information decrease is not a problem, because, while tracing steps the common interests among users also decrease.

It is difficult for the client-server architecture to emulate word-of-mouth style, because the central server gathers information from group members and restricts server access rights to only group members. This sometimes means that the friend of a group member might not be included in the group and will not be able to obtain information from the central server. On the other hand, P2P architecture that uses the word-of-mouth style traces the steps from friend to friend and is less likely to exclude people.

Approach: Emulate Word-of-mouth Information Distribution



I.U :Information Units that includes URLs of Web Sites.

2.2 Approach

Our goal is that a user can look up web sites interesting to him by getting URLs of web sites his friend previewed. We define as things that information that includes the URLs of web sites and titles of web sites seen by friends with the same interests. The reason to define the word called “information units” is that we think to extend information units to path information of the recently used on Windows OS and path of mail on a specific mailer (Sec 4.1). An information unit includes a URL of web site. In general, some information units, which include some URLs, are send with a transmission from a sender to a receiver. The figure illustrates how the information units are delivered in a word-of-mouth manner.

First, User A, the original content discoverer, who found relevant information on a Web server, opens and supplies information units, which include the web site URL he has previewed, to members registered on his friend list. In this case, User A’s friend list includes User B. Information units that include the URLs of User A’s web sites are sent from User A’s PC to User B’s PC. User B can access the web server contents discovered by User A using the URLs included in the sent information units. If User B accesses the content, we propose a mechanism in which information units supplied by User B include the URL of the content. User C, a friend of User B, can receive information units because he is included in User B’s friend list. Then, User C can also view the contents discovered by User A, because the information units are transmitted from User A to User B and User B to User C, the the same as a word-of-mouth transmission method. If User C views content using the information units received from User B, User D can also receive information units. This way, information units can be transmitted to a friend of a friend, even if he is not known by the content discoverer. User A sends the information units only to User B, but User A can potentially transmit the information units to User C and User D using word-of-mouth.

Issues to Develop a System

Most important issue is

- 1) Distributed Information must be relevant for receivers.

Other Issues are

- 2) Little effort for people to share information, and
- 3) The ratio of relevant information distributed
(number of information relevant for receivers
/ number of information distributed by our system)
should be high.

3.Issues

Developing a management system that emulates word-of-mouth transmission involves three important issues:

First and most importantly, the distributed information must be relevant to the receivers. If the management system distributes an information unit that has no value to the receivers, the information system has also no value. If the receiver can use the information effectively, then the distributed information has value. We confirmed the value of the distribution information in an experiment.

Second, people need to exert little effort to distribute information using word-of-mouth. Specifically, when a person finds a good Web site and introduces the web site to his friend, the person usually sends the URL by E-mail. Our management system minimizes effort required by the information supplier. Furthermore, our system enables the information supplier to supply information using only a web browser as usual, without additional actions.

Third, the ratio of relevant information distributed is high, because our system enables receivers to find useful information in relatively short time, because they need search the received titles and web sites URLs. If only small amount of relevant distribution information is available, the receiver needs much more time to find information. If the rate is high, the receiver can find information quickly and effectively.

System Design

- Emulate word-of-mouth by using P2P Architecture.
- Distribute Information units to access contents in the Internet by using WWW Browser.
- Target Web Browser is the Internet Explorer(IE).
- Extend distribution target to File Systems and Mails addition to the Internet contents.
- Distributed information is customizable by setting a configuration file in order to prevent confidential information leak.
- Our platform utilizes an existing P2P platform(JXTA) and an open source RSS Viewer.

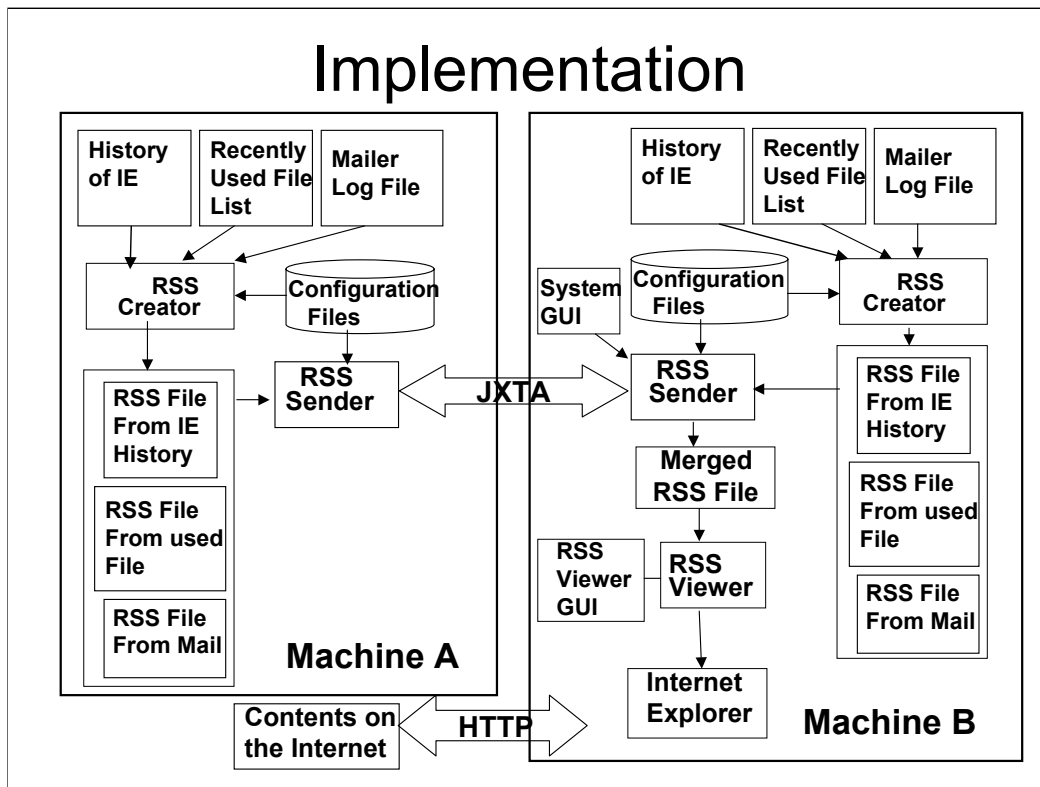
4. Development

4.1 System Design

Here, we introduces the development approach of our management system. Our management system features a P2P architecture that enables information units to be distributed from one friend to other friends by word-of-mouth. It distributes information units to access Internet contents using a web browser in RDF Site Summary (RSS) [6] form. Because structure of access histories depend on web browser, we used the Internet Explorer (IE), which is the most commonly used standard browser. To confirm extensibility, we extended the distribution target of our system from information units used to access Internet contents to include the path information of the recently used on Windows OS^{*1}, and the path of a mail on a specific mailer called 'Becky!' Internet mail [7]. The RSS file created from the mailer log includes a subject, a sender, whether there is an attachment file exists, and, if so, the name of attachment. To prevent confidential information from being leaked, and to filter irrelevant information, our management system had a mechanism to customize which information units are included in distributing target by setting configuration files.

Our system uses a P2P foundation software called JXTA[8], which enables general information exchange and an open RSS Viewer, called RSS Owl [9], in order to obtain reliability and reduce a development costs of our system. It is especially easy to determine if the client is using an RSS viewer by reconstructing the source code of of the RSS Owl 0.8b. Concretely, when RSS Owl reads the file and the user clicks an Internet title, RSS Owl accesses the content via the proxy server specified in IE before reconstruction. On the other hand, after reconstruction, the user can specify a proxy server explicitly from the GUI of the RSS Owl. This reconstruction enables a system administrator to distinguish between normal web access using IE and access using the RSS Viewer with distributed information. The system administrator can analyze frequency of using the distributed information unit by this reconstruction.

*1: We confirmed this movement test using Windows XP only.



4.2 Implementation

The word-of-mouth information management system uses P2P architecture. This figure shows that machine B is the PC that receives the information units and machine A is the PC that supplies the information units. Our management system's architecture can consist of three or more PCs. Machine B also can supply information units.

First, the RSS Creator reads the IE history, the recently used file list, and the mailer log file and makes three RSS files after reading the configuration file. It then reads the IE access history for the specified number of days given in the configuration file and creates an information unit including a web site URL of a Web site from each access history. Then it outputs an RSS file which consists of many information units and includes many web site URLs. The creator reads the paths of the recently used file list from a specified directory and creates information units from a recently used Windows OS files before, outputting an RSS file, based on the history. The creator reads the mailer log file and creates information units, then outputs an RSS file from Mail.

Second, machine B's system GUI displayed host names which our management system is activated. By specifying hostnames from the system GUI machine B's user can request to get information. Then, machine B's RSS sender asks machine A's RSS sender to send RSS files using JXTA. Machine A's RSS sender looks up the configuration file and confirms that it can supply the files to machine B, and returns three files. Machine B's RSS sender adds machine A's hostname to each unit in the three received RSS files and identifies the original host before, merging the three files to a merged RSS file. If Machine B's RSS sender received also requests RSS files from other machines and adds the sender's host name. If machine B's RSS sender received information units that accidentally includes the same web site URLs from other hosts, his machine merges them into an information unit and adds the other hostnames. This second step, Machine A's user supply information to use Machine B's user. To get information, Machine B's user need to only specify hostname and Machine A's user does not do anything. This results little effort to share information and solve the second issue.

Thirdly, the RSS viewer reads the merged RSS file from the GUI of the RSS Viewer. Machine B's user clicks a title of the web site. Then, the IE starts up in an RSS Viewer window and the user can see the contents of the specified web site. In this process, because the user views the content using IE, it is automatically recorded in the IE History and read during the RSS creator's next file creation process. If machine B is requested RSS files from host C, machine B can also supply the information units which were viewed using with RSS viewer, even if the information units were received from machine A. This process enables information transmission by word-of-mouth, without an additional operations.

Configuration Files

Configuration File example to create RSS file

[Local]

Interval=5

[Default]

Target-peer=*

Title=*

Access-span=7days

Path=!C:

Configuration File to access control

[Deliverer Settings]

Interval = 2

[Remote Peer]

Allowed Peer = hostA, hostB, hostC, hostD

4.3 Configuration Files

The configuration files can customize information to distribute and an access control list.

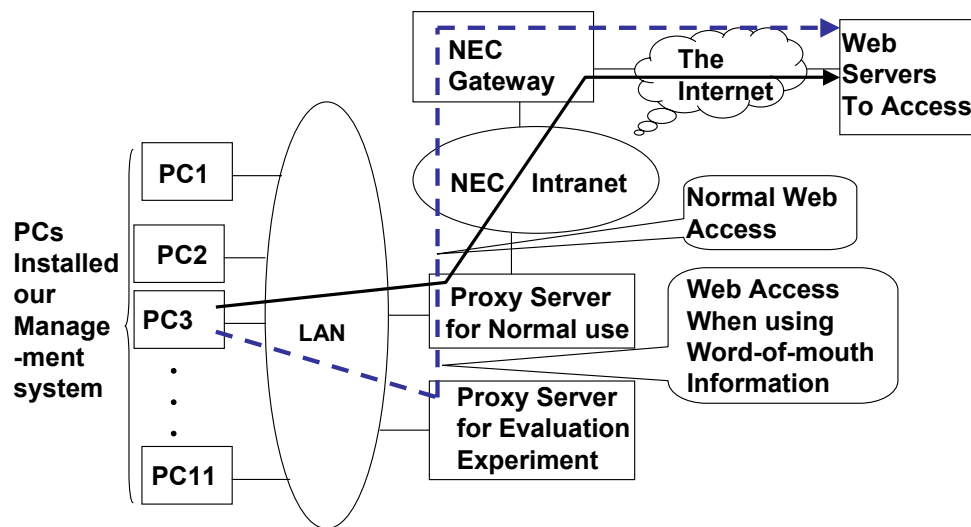
The upper figure is a configuration file for creating RSS files. This file specifies the filtering conditions included in the RSS files when the creator reads the IE history file, the recently used file paths and a log file of specific mail software to create RSS files. It also specifies the interval for creating the RSS files.

In other word

- 1) The interval line specifies the RSS file creation interval in minutes. This example specifies an interval of 5 minutes.
- 2) The target-peer line specifies which hosts are to include following a preset protocol. This example specifies the application of requests from all hosts is applied.
- 3) The title line specifies which titles include in the RSS file or which titles to exclude. This example specifies that all titles be included in the RSS files.
- 4) The access-span line sets age restriction on the information to be included in our system's RSS files. Here, we specified the last 7 days.
- 5) The PATH line specifies that RSS files include only specified paths or URLs, or to exclude specified paths or URLs by using the symbol "! ". Here, we excluded information from C: drive. Because local drive (C:) includes confidential documents in higher possibility, this drive is removed from the information distribution target to prevent confidential information leak.

The lower figure show an access control configuration. The interval line specifies the interval for discovering other hosts that run using the word-of-mouth information management system. Here, it is set for every 2 minutes. The allowed peer line specifies the host names that can reply to an RSS file request. This example specifies that the host replies to requests whose host name matches hostA, hostB, hostC, or hostD.

Evaluation Environment and Result



167 Web access occurred using our management system, (6 use/ person, day)

5. Evaluation

5.1 Evaluation experiment

We evaluated our word-of-mouth information management system in a LAN environment. If the person usually accessed web servers using IE from a PC installed our management system, he accessed the web server via a proxy server for normal use, NEC Intranet, or NEC Gateway. On the other hand, if the person accessed the web server using the RSS Viewer and received information units from another person, the access route would first be the proxy server for evaluation experience, and then the same route that usually used. By accumulating the proxy server for evaluation experiment's access logs, we measured how many times the information units were distributed using our system.

The configuration file for creating the RSS files used in this experiment is different from the sample configuration file described in sec. 4.3. Using only the Path line. In this experiment, the Path line was is "Path=!C://D://E://.nec.co.jp/". Because local drives (C://D://E://) includes confidential documents, there is higher possibility that, these drives will be removed from the information distribution target. Because web sites in the NEC intranet (.nec.co.jp) are viewed by many NEC people, they required less merit to share, compared to other sites, and intranet sites were filtered. The configuration files for access control were set by all the people who participated in this experiment.

Our management system was used by 9 people over 5 days and resulted in 167 accesses to Internet contents using distributed information. The total number of participant in the 5 days period were 28 and the average amount of information distributed by our management system was used 6 items per day per person.

Discussion

- Evaluation result shows that our management system can transfer 6 valuable information per day per person. This means distributed information includes valuable one for a receiver. (1st Issue)
- Information receiver specifies hostnames which want to get information from GUI, and information supplier only activates the system and does not do anything.(2nd Issue)
- 4% pointer information is used from distributed information. (3rd Issue has a little problem)

5.2 Discussion

On the first day, 9 people joined the experiment and generated 63 Internet using distributed information and our word-of-mouth information management system. However, from the second to fifth day, the number of users decreased to about 5. The reason is that all the distributed information units were new to the receivers on the first day. But the distributed information on the second day included the same information on the first day. Because the distributed information units only on the second day were restricted to the information users access from the first day to the second day, there were fewer new information units on the second day. This caused the decrease in the number of users.

Information distributed using our management system was used 6 times per day per person. This information utilization means that distributed information must include sufficient useful information for the receivers. Our management system, therefore, solves the first issue.

A characteristic of our management systems is that the information receiver only to specify hostnames which want to get information from our system GUI, and information supplier only activates the system and does not do anything to supply information. This solves the second issue.

I explained the rate of distributed information used by a receiver in the evaluation experiment. The average number of web accesses outside the NEC intranet, which is not filtered out by the configuration file for creating RSS files, is about 70 times per day per person. Because our system obtained only recent history from previous seven days, on the first day we averaged 70 times per day per person multiplied by five weekdays and multiply 9 people, resulting in 3,150 pieces of information distributed. After the second day, only adding new information to the first day's information resulted in 4,480 items. Because the configuration file for creating RSS files filtered out the local drive and no one used the mailer, there were few non-web items and we can ignore them. The rate of distributed information units used by receiver was about 4%, which we calculated by dividing 167 into 4,480.

For the third issue, the ratio of distributed information units used by receiver was only 4%. There is possibility that the distributed information in our management system includes a lot of irrelevant information. If our system had a mechanism to remove clearly noisy information for receiver automatically and doesn't send these noisy information, we could think that the rate of distributed information used by receiver become higher. Our system needs a improvement to increase the rate.

Related Work

- Poblano [Yeager 2002]: an example of P2P application using JXTA. Ranks contents by using reputation in a group. But this needs a keyword to search contents.

Our platform needs **no keyword**.

- del.icio.us: a bookmark sharing tool.

A user needs to add bookmarks explicitly to change sharing information.

Our platform can increase sharing information by **only using a Web Browser naturally**.

6. Related Work

Poblano[10] proposed an application system using JXTA. This application sorts web content URLs based on keyword specified by a user. To achieve this, Poblano proposed using a special interest group who share interests in a specific category, based on the reputations of the members, who can search a specific member files for content using keywords. It then sorts the contents in order from the top, which is thought to have the most relevance to the user. The author's reputation rises when a document written by the author is looked at frequently by group members. This can result in the content or author being ranked higher in the group. Poblano manages these reputations.

Because Poblano needs adequate keywords to obtain content, users cannot find potentially interesting content unless they can specify enough keywords. On the other hand, because our word-of-mouth information management system can transfer information units to access contents without any keywords, it is superior to Poblano from the perspective of enabling access to interesting content.

Del.icio.us[11] is a management and sharing tool for web browser bookmarks. This tool organizes bookmarks by the most recent date used and enables them to be shared with others. It enables web site access to bookmarks registered using in bookmarks of other.

To change sharing bookmarks, del.icio.us users need an explicit operation to add web sites to a bookmark. Their bookmarks does not change without explicit operation, while our management system gets distribution information from an IE history file, making it superior to the del.icio.us by enabling to change the amount of distributing information by only using IE and the web site history.

Conclusions and Future Work

- Information heard by word-of-mouth is valuable and trustworthy.
- We developed a word-of-mouth information management system by using P2P platform JXTA and aimed to solve 3 Issues, 1) Distributed Information must be relevant for receivers, 2) Little effort for people to share information, 3) The ratio of relevant information distributed should be high.
- Evaluation result shows that this platform can transfer 6 relevant pointer information per day per person (1st Issue).
- Information receiver specifies hostnames which want to get information from GUI, and information supplier only activates the system(2nd Issue)
- A future work is to eliminate irrelevant information from distributed information to reduce time to find valuable information (3rd Issue).

[Acknowledgement]

This research was sponsored by the Ministry of internal Affairs and Communication in Japan.

7. Conclusions and Future Work

The amount of information available over the Internet is huge, and it is difficult and time consuming to find relevant actual information. The information obtained from friends with the same interests is usually valuable and trustworthy. Web sites accessed by community members with the same interests can include a greater proportion of interesting information than ranked web sites which the member can search using a keyword search engine. We developed an information management system that emulates word-of-mouth style, and aimed to resolve three issues, 1) Distributed Information must be relevant for receivers, 2) Little effort for people to share information, 3) The ratio of relevant information distributed should be high.

To solve these issues, we implemented a system that enables a URL exchange between people with the same interests by obtaining URLs from a recent IE history file. In our evaluation experiment, the total number of people in 5 days was 28 persons, and the information distributed by our management system included 6 relevant information unit per day per person to access useful web sites. Because 6 times per day per person confirms that the distributed information includes sufficient useful information for the receiver, our management system resolves the first issue. A characteristic of our systems is that the information receiver only to specify hostname which want to get information from our system GUI, and information supplier only activates the system and does not do anything to supply information. This solves the second issue. For the third issue, the rate of distributed information used by the receiver was only 4%. There is possibility that that our management system includes a lot of irrelevant information. In the future, we plan to eliminate this irrelevant information from the distributed information to reduce the time for the receiver to find valuable.

[Acknowledgement]

This research was sponsored by the Ministry of Internal Affairs and Communication in Japan.

[References]

- [1] Ivan Misner, "Word-of-Mouth: The Word's Best-Known Marketing Secret", <http://www.entrepreneur.com/article/0,4621,301179,00.html>
- [2] Shen, H.T, Shu, Y, Yu, B., "Efficient semantic-based content search in P2P network", Knowledge and Data Engineering, IEEE Transactions on Volume 16, Issue 7, 2004
- [3] Lechner, U., Schmid, B.F., "Communities-business models and system architectures: the blue print of MP3.com, Napster and Gnutella revisited", Proc. of 34th annual Hawaii International Conference on System Sciences, 2001
- [4] Ripeanu, M., "Peer-to-peer architecture case study: Gnutella network", Proc. of First International Conference on Peer-to-Peer computing, 2001
- [5] Salman A. Baset and Henning Schulzrinne, "An Analysis of the Skype Peer-to-Peer Internet Telephony Protocol", <http://www1.cs.columbia.edu/~library/TR-repository/reports/reports-2004/cucs-039-04.pdf>, 2004
- [6] W3C, Resource Description Framework (RDF), <http://www.w3.org/RDF/>
- [7] RimArts, Inc., "Becky! Internet Mail", <http://www.rimarts.co.jp/index.html>
- [8] Yeager, W., Williams, J., "Secure peer-to-peer networking: the JXTA example", IT Professional, Volume 4, Issue 2, March-April, 2002
- [9] "RSS Owl!", <http://www.rssowl.org/>
- [10] Rita Chen and William Yeager, "Poblano A Distributed Model for Peer to Peer Networks", <http://www.jxta.org/docs/trust.pdf>
- [11] "del.icio.us", <http://del.icio.us/doc/about>