A Proposal of an Adaptive P2P Database Retrieval System Using a Hierarchically Categorized DHT Structure

Takuya Sasaki*, Hideyuki Akasaki*, Jun Sawamoto*, Takashi Katoh*, Yuji Wada**, Norihisa Segawa* and Eiji Sugino*

* Faculty of Software and information Science, Iwate Prefectural University, Japan
** Department of Information Environment, Tokyo Denki University, Japan
g031e078@edu.soft.iwate-pu.ac.jp, sawamoto@iwate-pu.ac.jp

Abstract –In recent years, there are various information sources in our surroundings in accordance with the development of ubiquitous equipments and information and communication technology. However, as for the management of those information sources, it is the usual case that individual and independent database management system is used for managing and maintaining each information source. In this paper, we propose an adaptive database search system GrowApp which unifies multiple heterogeneous DB's in ubiquitous Peer-to-Peer (P2P) network environment. In GrowApp, we apply the mobile agent technology in DB search mechanism and build a virtual DB environment integrating multiple heterogeneous DB's in P2P network system. We introduce hierarchical grouping structure into DHT in P2P and enhance conventional keyword search capability. In addition, we consider mechanism of automatic generation and reformation of grouping structure based on the users' retrieval records. Finally, we discuss implementation and evaluation of the system.

Keywords: P2P network, virtual database, mobile agent, database search, data mining.

1 INTRODUCTION

In recent years, as the ubiquitous environment spreads, various services are requested to be provided according to the place (location) and time, etc. for the user who acts while moving around with a mobile device. The free mobile computing environment with an ad hoc communication environment like wireless LAN and the short distance wireless telecommunications, etc. is coming to be achieved and removing the restraint on user's movement [1].

On the other hand, it is said that the technology that enables the discovery and integration of dynamic services is indispensable as the mechanism that provides ubiquitous services. It is requested to adjust to user's favor, and the change in the context including location, time and user's situation dynamically. The application of technologies of semantic Web and the Web service, etc. is thought as a mechanism that dynamic services corresponding to such ubiquitous environment are achieved [2]. However, the problem of the management of information sources remains[3].

In ubiquitous environment, there are various information sources in our surroundings in accordance with the development of ubiquitous equipments and information and communication technology. It becomes possible for data sources distributed in various places to tie to the data processing nodes along with the existence of the ubiquitous network environment. However, as for the management of those information sources, it is the usual case that individual and independent database management system is used for managing and maintaining each information source.

In this paper, we propose an adaptive database retrieval system GrowApp which unifies multiple heterogeneous DB's in ubiquitous Peer-to-Peer (P2P) network environment. In GrowApp, we apply the mobile agent technology in DB retrieval mechanism and build a virtual DB environment integrating multiple heterogeneous DB's in P2P network system.

We introduce hierarchical categorization (grouping) structure into distributed hush table (DHT) in P2P and enhance conventional keyword search capability [4, 5, 6, 7]. Individual data bases are integrated by using the virtual schema [9], and all the database searches are enabled with scanning each database by the agent. Integrated databases are viewed as one P2P node and registered to the DHT and managed. Moreover, retrieval by category is enabled by using a hierarchically categorized DHT structure [8].

In addition, we consider mechanism of automatic generation and reformation of grouping structure based on the users' retrieval records. In the DHT, it is a burden for the network manager to renew the hush table when a new group appears or an old group is deleted. Then, the automatic grouping is examined in this paper. When a user accesses the system and retrieves data, we can consider a set of continuous data retrievals as a candidate for grouping. When not only one user but also multiple users are doing a similar set of retrievals, they could be registered in the DHT as a new group. At the same time, the group is updated in proportion to the retrieval frequency. Less frequently retrieved groups are deleted from the table. Thus, the user can find favorable data more easily and efficiently.

Finally, we discuss implementation and evaluation of the system. In the present study, it is targeted to simulate and to evaluate whether the automatic grouping of the DHT is carried out in P2P environment. The Kademlia algorithm [6] is considered appropriate for large number of nodes and is used in this simulation.

2 GrowApp SYSTEM

2.1 Features

GrowApp is an intellectual search engine intended for the user who uses the mobile communication environment. It aims to retrieve information that exists in all places of the society.

GrowApp is not a centralized database management system. In recent years, the number of ubiquitous databases which are managed and maintained by the individual, the group, and the store, etc. is increasing because of the development of the network system. GrowApp aims to combine those existing databases and to offer profitable information for the user.

The flow of the system until the user, using the mobile communication terminal, retrieves information, and acquires it is as follows.

(1) The user accesses GrowApp.
(2) The user inputs keywords to the system for retrieval.
(3) The system that acquires the keywords, generates the retrieval agent who possesses user's context and DHT.
(4) The retrieval agent looks up DHT with the hashed values of the user's retrieval keywords.
(5) When the node with contents for the value of the retrieval keywords is found, the retrieval agent moves to the node.
(6) The retrieval keywords are thrown to the virtual database management agent at the node with contents.
(7) The virtual database management agent who receives the retrieval keywords executes query to the virtual database at the node.
(8) The acquired value, by the execution of the query, is returned to the retrieval agent.
(9) The retrieval agent who acquires the value returns to the original user's position, and presents the acquired information to the user.

2.2 Structure

GrowApp is composed of four major elements of virtual database, mobile agents, the P2P network, and the DTH.

2.2.1 Virtual Database [9]

Various kinds of databases exist in the society. Because the management technique is different, they cannot adopt the same query method. To retrieve, and to update them from the application side, regardless of the kind of the database, the virtual database schema is set up between the user and databases. It converts the search queries and the virtual database schema doesn't ask the kind of the database. Figure 1 shows that the user accesses various databases without considering heterogeneity of each database by using the virtual database schema.

Various databases are assumed to be a retrieval object in GrowApp. It becomes a procedure of the database management agent's to execute the query to the virtual database schema and then the virtual database schema retrieves databases and acquires information.

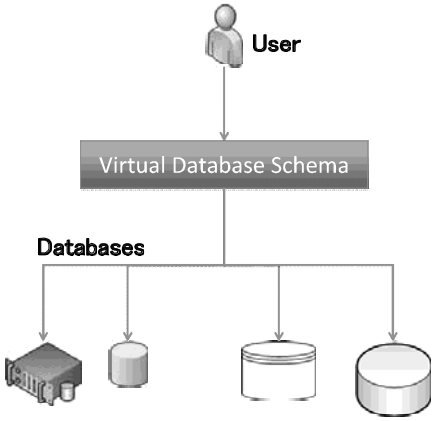

Figure 1: Structure of virtual database.

2.2.2 Mobile agent

A virtual database as stated in section 2.2.1 could include all the databases in society under its schema, forming a huge virtual database. To view all the databases under unified schema is ideal. However, we can easily see a problem of communication bottleneck between multiple users and the virtual database schema and it is hard to consider a unique virtual database schema.

In this paper, the virtualization is done by using the virtual database schema for a certain number of databases. The database management agent is arranged in each virtual database where those virtualizations were done.

In addition, when the user uses the system and the retrieval keyword is turned on, the system generates the retrieval agent who has user's context. It is assumed that user's context is the one to show the situation that relates to the user at the point of the retrieval. A past retrieval history can be easily reflected in a present retrieval by adding a past retrieval history in addition to spatial information and temporal information to the context.

When the retrieval agent retrieves data as shown in Figure 2, query is thrown out to the database management agent. The database management agent, who receives query, retrieves data from the heterogeneous databases by executing query for the virtual database schema.

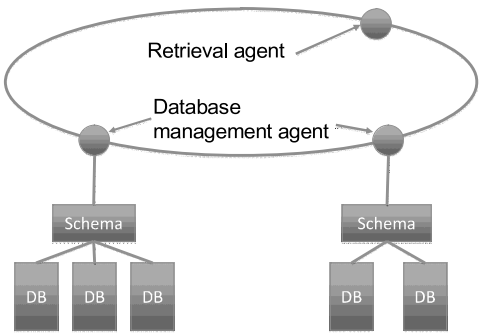

Figure 2: Retrieval agent and database management agent

2.2.3 P2P network and DHT

When the database is brought together in one place, the bottleneck between databases and users is caused. To solve

his problem, when the user retrieves data, a retrieval agent is generated, and information is acquired according to the P2P communication between agents.

Ad hoc participation of the nodes to the network can be expected. Then, assuming frequent arrival and departure of participating nodes, the DHT that uses Kademlia algorithm [6] is introduced, and the P2P network is constructed. As for the Kademlia algorithm, it is evaluated as comparatively easy to implement because special update work is unnecessary under such frequent updates condition.

2.3 Functions

GrowApp has the following functions as an intelligent search engine that considers the mobile computing environment intended for a lot of heterogeneous databases.

2.3.1 Current user location

The user is expected to participate in an ad hoc system in the mobile computing environment. Then, user's spatial information is acquired, and the spatial information is treated as user's context. The user's spatial context is used to understand the user's situation and to put the weight to most suitable information.

When the user accesses the system, because the GPS function is installed in today's cellular phone, user's GPS information can be acquired. This information is acquired, and maintained by the generated retrieval agent of each user.

2.3.2 Grouping of contents in DHT

This system constructs the P2P network. When the node retrieves contents, the DHT is used. Retrieved contents are downloaded by tracing down the node that maintains contents from the query that the user executed referring to the DHT. However, the perfect matching with the search keyword is demanded in the hush table. The DHT has the drawbacks that the partial matching or the group matching retrieval cannot be achieved. It is difficult to do a flexible retrieval. This becomes a problem that should be solved when the search engine is constructed.

Then, the DHT is hierarchically categorized, and contents are grouped. Flexibility can be introduced into the user's retrieval by providing information with high possibility to be retrieved at the same time to the search keyword by this contents grouping. It is thought that it can be supplemented to lack of a flexible retrieval function that is the weak point of the DHT.

2.3.3 Dynamic computation of category by data mining

When the user retrieves many of the databases that exist in the society, a large amount of information will be presented to the user. It is very difficult to discover profitable information for the user [10].

Then, the category is measured by data mining. The retrieval agent maintains a past retrieval history of the user besides user's positional information. When the user uses the system and retrieves data, the preference of user's retrieval history and acquired information is measured. For the measurement method, the system puts higher priority for contents the user downloaded in the past retrieval. It becomes possible to present information that is thought to be a reflection of the result of the measurement and effective to search preferred data from among a lot of retrieval results.

3 AUTOMATIC GROUPING IN DHT

3.1 Outline

Generally, the cost to renew the hush table in DHT is high. Churning of the node occurs frequently because ad hoc participation and withdrawal of node is assumed. In addition, the frequency in which the DHT is renewed increases if it comes to group contents according to user's retrieval results frequently. Then, we propose an automatic grouping function of the DHT. It becomes possible to offer contents grouping and renewal of DHT in real time by this function.

In this paper, the Kademlia algorithm is used for the DHT. As for the Kademlia algorithm, it is evaluated as a strong algorithm for the ad hoc node participation and secession. When a new category, in addition to participation and the secession of the node, was created and an old category is deleted the hush table is renewed. The accuracy of the contents retrieval decreases if DHT doesn't update according to the change of nodes. Though updating DHT in real time is ideal, the updating process requires user's work. Then, we propose a method of automatic generation of new contents group and deletion of old contents group according to user's retrieval situation.

It is possible to view those retrieval contents as one group when a certain user uses the system, and user retrieves not only one content but multiple contents continuously. The group of contents that a certain user used has the possibility to be used as a useful content group for other users. When not only the user but also another user is doing a similar retrieval, they could be registered in the DHT as a new category. However, when a flow of time and a new kind of item appear, the item that comes off as a retrieval target is not permanently left as a member of category but the category is updated in accordance with the retrieval frequency etc. It becomes possible to make the retrieval efficient and suitable for the preference that the user is always requesting.

3.2 Example of group creation and deletion (example of PC parts)

An example flow of the automatic creation of a new group that occurs when the user actually retrieves and contents are downloaded, and deletion of a group when it becomes not used, is described as follows.

(1) Contents registered in the DHT do not belong to a group in the initial state. Contents are shown in Figure 3. "P6T" is a name of a mother board, "CD-

552GA" is a CD drive and "Core i7 920" is a CPU respectively.

(2) These three contents are retrieved and downloaded frequently by many users.
(3) These three contents are registered in the DHT as a group retrieved often at the same time. Grouped contents are shown in Figure 4.
(4) DVD drive appears. The DVD drive is a content named "DVR-216DBK" here. The state is shown in Figure 5.
(5) The user comes to do the retrieval of the combination of DVD drive, CPU, and mother boards by the appearance of the DVD drive in place of the group that contains the CD drive.
(6) The group is newly generated because the combination of DVD drive, CPU, and the mother board frequently came to be retrieved by the user, and it is registered to the DHT. Figure 6 shows the state in which a group is newly generated.
(7) The group including the DVD drive comes to be retrieved and downloaded by more users than the group of the CD drive. The CD drive after the DVD drive appears is retrieved and downloaded less frequently.
(8) The group that contains the CD drive enters the state that is with less profitable information for the user. Then, the group that contains the CD drive is deleted as shown in Figure 7. However, only the group of the CD drive is deleted, and CD drive itself is possible to be retrieved and downloaded as a content.

Figure 3: Initial state without a group

Figure 4: A group is generated

Figure 5: Appearance of DVD drive

Figure 6: New contents group is generated

Figure 7: The group that contains the CD drive is deleted

4 SYSTEM IMPLEMENTATION

The system consists of the User Interface part, mobile agent, P2P network, and the virtual database. Programming language Python is used for implementation. The function of contents grouping in DHT is implemented on the P2P simulator with Kademlia algorithm, and the usability is experimented and evaluated. Then, we continue to build the virtual database schema, the retrieval process, the database management agent and the user interface of GrowApp.

User Interface has the function to display information to the user and the form that input the retrieval keywords. This system uses two kinds of mobile agents. One is a retrieval agent who receives user's retrieval keyword and starts retrieval of contents information. And, the other is a database management agent who receives contents information in the virtual database and query from the retrieval agent. For the virtual database, multiple databases are virtualized by using the virtual database schema prototyped by the project [9].

5 EVALUATION

Here, we measure the communication traffic to estimate the degree of burden to the user to see whether the automatic grouping of contents in the DHT is useful or not. And at the same time, retrieval performance is evaluated quantitatively.

5.1 Outline of the evaluation

Outline of the evaluation is shown as follows. Number of PCs (install about 200 nodes per PC) is increased and the following parameters are measured.

- Measurement of traffic volume of information
- Investigation of retrieval performance (precision, recall)
- Response time of retrieval

In the evaluation, about 200 nodes are generated per one PC, and P2P network is constructed. And the measurement of traffic is carried out and retrieval performance with the DHT is verified.

5.2 System structure

Figure 8 shows the system structure of the evaluation. In the evaluation, in order to evaluate the accuracy of the DHT with automatic grouping and the communication traffic, we generate more than 200 nodes and construct simulated P2P network. Whether contents in the DHT are added accurately as a contents group based on user's retrieval keyword, and how communication traffic changes compared with the client-server model are verified.

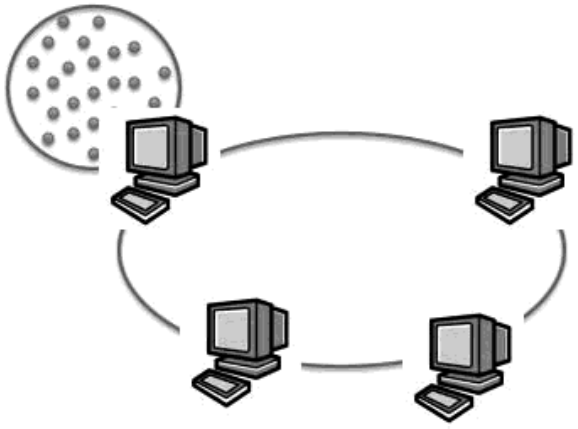

Figure 8: System structure of the experiment

5.3 Software configuration

Figure 9 shows the software configuration. The P2P simulator is started on each terminal. As use of various mobile communication terminals are assumed, both Windows and Linux are used for OS of experiment. Put the hush operation on the retrieval keyword from the user and refer to the DHT for the P2P simulator. A mobile agent moves to the specified node position, and the mobile agent communicates with the database management agent, and demands and acquires contents information that the user is requesting. Contents information is displayed to the user at the user's interface.

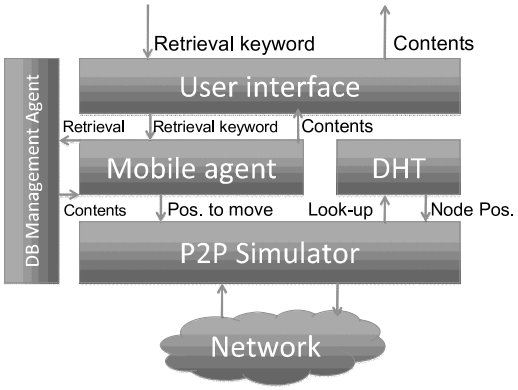

Figure 9: Software configuration

5.4 Evaluation

Items of the evaluation corresponding to the content of the experiment are as follows. The client-server model that stores the same volume of information as the experiment system in a database is used for comparison.

- Comparison of communication traffic when the experiment system and client-server model carry out the same retrieval operation.
- Accuracy evaluation of contents that are retrieved with keyword search.
- Confirmation of automatic grouping of contents in the DHT when multiple contents are retrieved simultaneously.
- Comparison of response time when contents are retrieved by this system and client-server model.
- Evaluation of the system under the condition of increasing number of participating nodes.

Based on the results of the experiment, the contents grouping in DHT is evaluated whether a useful measure or not. Evaluation criteria are communication traffic for DHT construction, retrieval, update, comparison with client-server model, retrieval performance for the user request. As a result of the actual experiment, the function of contents grouping of the DHT will be improved in the next step. Then, the implementation as a prime function of GrowApp will be examined.

6 CONCLUSION

In this paper, we proposed an adaptive and intelligent information retrieval system GrowApp targeting existing databases in the ubiquitous network society. We intend to implement and evaluate the GrowApp's primary function of grouping of contents in DHT. Based on the evaluation, improving the function of grouping of contents, the interface part with the virtual database schema, mobile agent, and user interface part will be integrated, and it is scheduled to go finally the system design of the entire GrowApp. In addition, the verification of the usability is scheduled to be covered in consideration of a concrete application system of the proposed GrowApp, and experimenting with the application.

ACKNOWLEDGMENT

This work was supported by Grant-in-Aid for Scientific Research (C) (20500095).

REFERENCES

[1] D. Chakraborty et al., "Toward Distributed Service Discovery in Pervasive Computing Environments," IEEE Trans. Mobile Computing, vol. 5, no. 2, pp. 97–112 (2006).

[2] Hiroyuki OGURA, Saeko MURAKAMI, Hiroyuki SATO, Tomihiko KOJIMA, Noboru SHIMIZU, Itaru HOSOMI, Application Systems on Semantic Web, Journal of Information Processing Society of Japan 43(7) pp.742-750 (2002).

[3]Ryan Huebsch, Brent Chun, Joseph M. Hellerstein, Boon Thau Loo, Petros Maniatis, Timothy Roscoe, Scott Shenker, Ion Stoica, and Aydan R. Yumerefendi. The architecture of pier: An internet-scale query processor. In The Second Biennial Conference on Innovative Data Systems Research (CIDR), Asilomar, CA, January (2005).

[4]Ben Y. Zhao, John Kubiatowicz, and Anthony Joseph. Tapestry: an infrastructure for fault-tolerant wide-area location and routing. Technical Report UCB/CSD-01-1141, U.C. Berkeley, April (2001).

[5] Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek, and Hari Balakrish-nan. Chord: A scalable peer-to-peer lookup service for internet applications. In Proceedings of the ACM SIGCOMM '01 Conference, San Diego, California, August (2001).

[6]Petar Maymounkov, David Mazi`eres, Kademlia: A Peer-to-peer Information System Based on the XOR Metric, In Proceedings of the 1st International Workshop on Peer-to-Peer Systems (IPTPS), March (2002).

[7] HIROKI KUZUNO, YUSHI NAKAI, ATSUMU WATANABE, TAKUYA KAWAHARA, TAKASHI KATOH, BISTA BHED BAHADUR, TOYOO TAKATA, A Proposal of Distributed Internet Monitoring System Using Mobile Agent, Transactions of Information Processing Society of Japan 47(5) pp.1393-1405 (2006).

[8] Yi WAN, Takuya ASAKA and Tatsuro TAKAHASHI, A Hybrid P2P Overlay Network for Non-strictly Hierarchically Categorized Content, IEICE Transactions on Communications E91-B(11):3608-3616 (2008).

[9] Y. Watanabe, K. Shoubu, H. Miita, Y. Wada, J. Sawamoto, T. Katoh, Virtualization technology for heterogeneous databases, FIT2009 (2009).

[10]K. Mori, S. Kurabayashi, N. Ishibashi, Y. Kiyoki, An active in formation delivery method with dynamic computation of user's information in mobile computing environments, IEICE, DEWS2004, March (2004).