6-2 Interconnection of Heterogeneous Databases by Correlation Measurement based on Semantic Space Model

ZETTSU Koji and NAKANISHI Takafumi

In this paper, we will introduce the new technology of correlation search a retrieval technique based on measurement space model. There are said to be two types of human's knowledge. One is the kind we obtain from reading literature. This may be called absolute knowledge. With ICT technology, it is knowledge that can be described by the ontology used in a conceptual dictionary and Semantic Web. Another kind may be called relative knowledge. This is paradigmatically, relational knowledge we gain when we compare knowledge A with knowledge B. Such knowledge may be connected in thought without being displayed (i.e, as a document on the Internet or being inscribed on a piece of paper). Humans often compare heterogeneous things. Moreover, humans are thought to unconsciously create the criteria for linking heterogeneous things. In this model, the criteria are represented by a set of axes which form space and by a measurement method which mathematically measures norms, distance or inner product. It is important to search not only for the correlation between heterogeneous knowledge, but also for factors (axis, index) that contribute to the correlation. This model will show correlations between heterogeneous knowledge and their factors, by deriving a selected knowledge cluster that has high correlation magnitude and a set of axes used to calculate the magnitude.

Keywords
Interconnection of heterogeneous knowledge bases, Link-free browsing, Knowledge GRID

1 Introduction

On the Web, information resources are growing rapidly. To access this information, Web search engines have been broadly used for some time. Though search engines enable us to search any page whose content matches the search query, each page in the search results is normally independent. So if a user wants, for instance, to do comprehensive research on a particular event or a particular concept, she has to check and arrange information according to your own interests. This becomes a more serious problem when you try to grasp, through information on the Internet, various events that occur in the real world. For example, because natural disasters influence a variety of domains, it is preferable to obtain information from a variety of sources. But the traditional search engine does not even give us access to relevant information, unless we input relevant terms. Thus, a mechanism of deriving suitable terms, from various domains, (in accordance with the objective and the intention of the part of the user) is called for.

With this background in mind, we are developing a method of integrating heterogeneous knowledge information sources *(knowledge bases)* in terms of their correlation and of searching for information across heterogeneous domains. This paper proposes a heterogeneous knowledge base integration method by correlation measurement based on semantic space model and outlines the system which

applies this method to Web browsing. In Section **2**, we will investigate relevant research and in Section **3**, we will explore the outline of proposed method. In Section **4**, we will show assessment experiment results and in Section **5**, we will explain an applied system. In Section **6**, we will give a summary.

2 Relevant research

Semantic Link Network (SLN) [6] is a general as well as flexible data model, which generates relational links among data based on a set of relational inference rules that constitute a semantic network. Our proposed method also forms the same semantic data model as SLN. It differs from SLN in deploying correlation measurement [15][16] in generating the relational links. In addition, we aim at integrating heterogeneous knowledge bases (from various domains) and acquiring concept terms from a variety of relevant domains. The results are represented as a semantic correlation network similar to SLN. In this semantic correlation network, a concept term is represented as a node and a correlation between two concept terms is represented as an edge. We also propose a data browsing method called *"Link-free Browsing"* which makes use of this semantic correlation network.

A lot of research has taken place with regards to auto-link generation. One sample from the literature [7] stresses the necessity for integration between searching (by way of information search system) and browsing system by way of hypertext system. The literature suggests that such integration is necessary especially in cases when a user is not sure about storage location of documents or when the number of documents exceeds the number of storage locations. To understand the criteria on the integration, see the discussion on Trellis system [9]. Web Watcher [10] learns the history of previous searches (by a user) and leverages it for Web navigation. Letizia [11] proposes a method of predicting Web pages that may attract user's interest by tracking the clickstreams and the history of past searches. Other literature [12] proposes the method of dynamically recommending links by classifying users who visit a Web site in accordance with their access patterns. This system proposes a method of correlation computation and representation across various domains, with the use of semantic correlation networks.

The traditional way of deriving (knowledge) information across different domains has been the bridge concept. Schema mapping [16] and bridge ontology [17] are its paradigmatic instances. Traditional method stresses defining a precise bridge concept in advance, which is a very difficult task. Moreover, the method is applicable to relatively minor problems and in the limited domain. We aim at integrating information across as many domains as possible by taking dynamic computing correlation approach, even if the accuracy is sacrificed.

3 Meta level integration of knowledge bases

Figure 1 illustrates a heterogeneous knowledge base access that we propose. To a user's query, it searches knowledge bases (from a variety of different domains on a Knowledge GRID) [10] for a set of highly correlated records. Search results are represented by semantic correlation network. Users are able to follow data from a certain knowledge base to data of another knowledge base along the semantic correlation network. In Knowledge GRID, a GRID node is set in each site and the knowledge bases are constructed on it. As shown in Fig. 2, on each GRID node, specialists in different domains meet together and cooperatively edit knowledge bases. From knowledge bases constructed on these GRID nodes, the system finds out correlating data in a variety of contexts and links these knowledge bases based on correlation.

3.1 Knowledge processing services

To bring to realization the framework shown in Fig. 2, we define two sorts of knowledge processing services. They are *intra oper-*

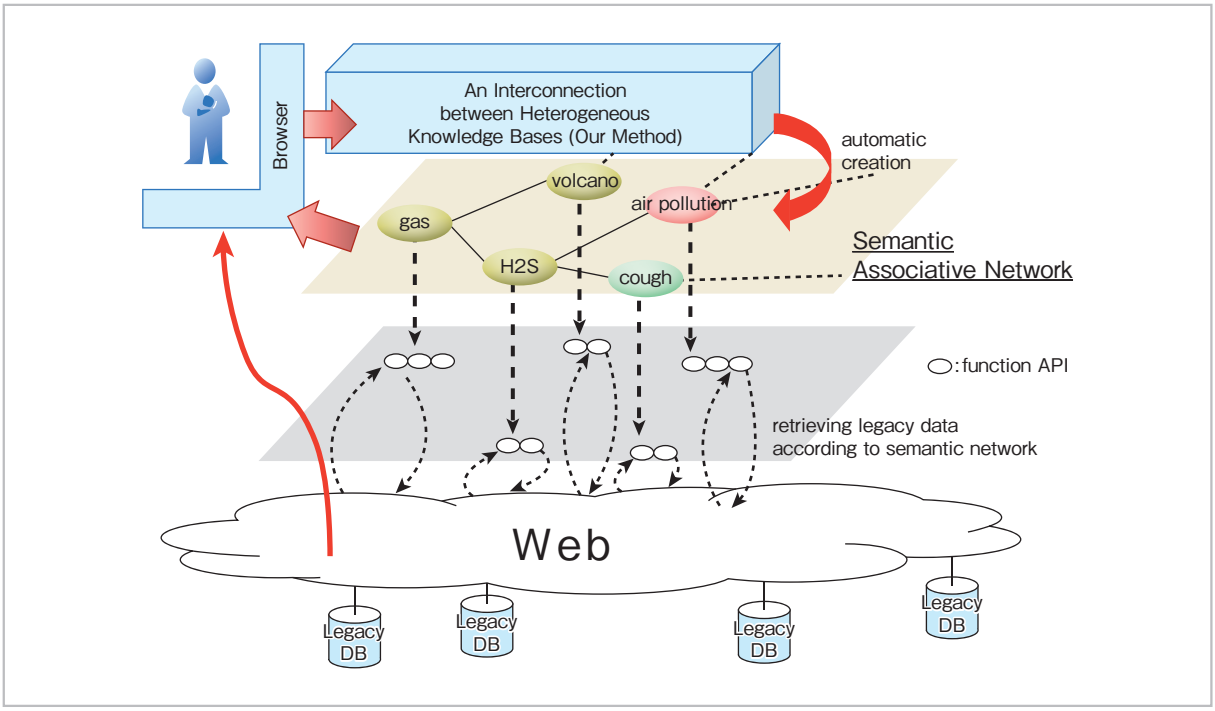

Fig.1 *Overview of heterogeneous knowledge base access by way of semantic correlation network*

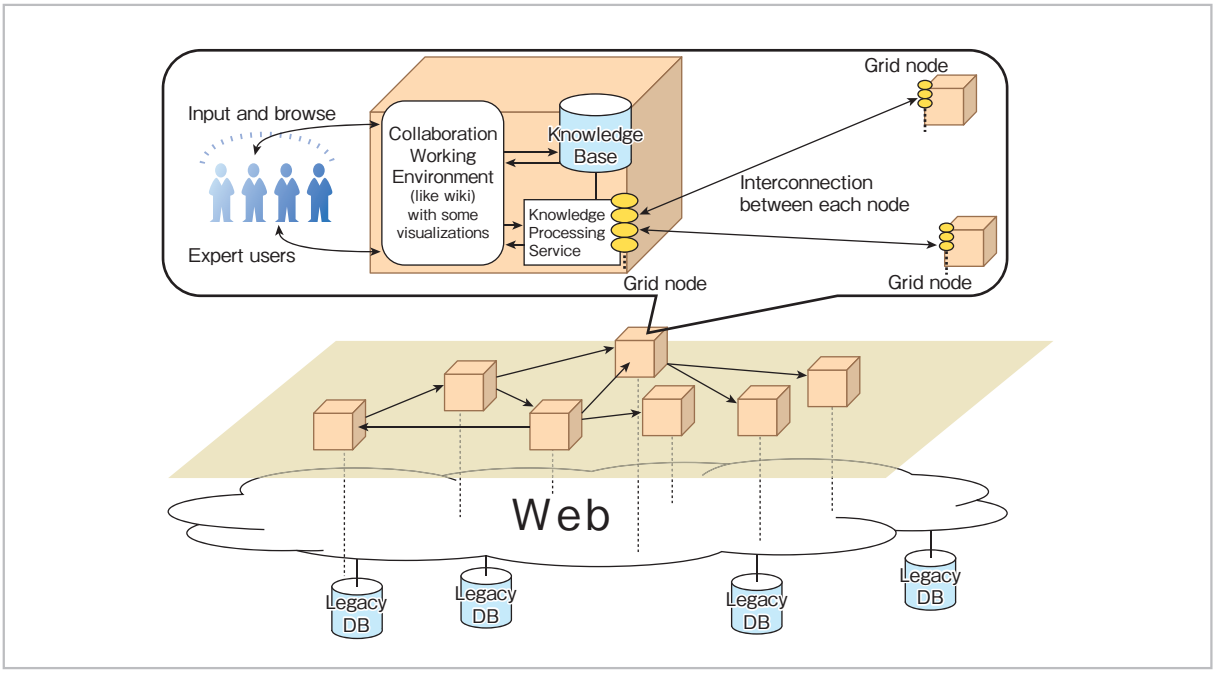

Fig.2 *Interconnection of heterogeneous knowledge bases on Knowledge GRID*

ation service and *inter operation service*.

Intra operation service: provides a function of correlation search within a single knowledge base. Given a set of concept terms as input, it searches within the knowledge base for a set of concept terms that are highly correlated with the input. It then outputs them and a value (correlation coefficient) that stands for the strength of correlation.

Inter operation service: provides a function of correlation search among multiple knowledge bases. Though input and output are the

same in intra operation service, inter operation service differs in that, a set of concept keywords which constitutes input and output belongs to different knowledge bases. Figure 3 represents the process of generating a semantic correlation network with the use of these two sorts of services. To understand more about this process, we will explain the procedure of combining two knowledge bases (termed knowledge base A and knowledge base B respectively) using Fig. 4.

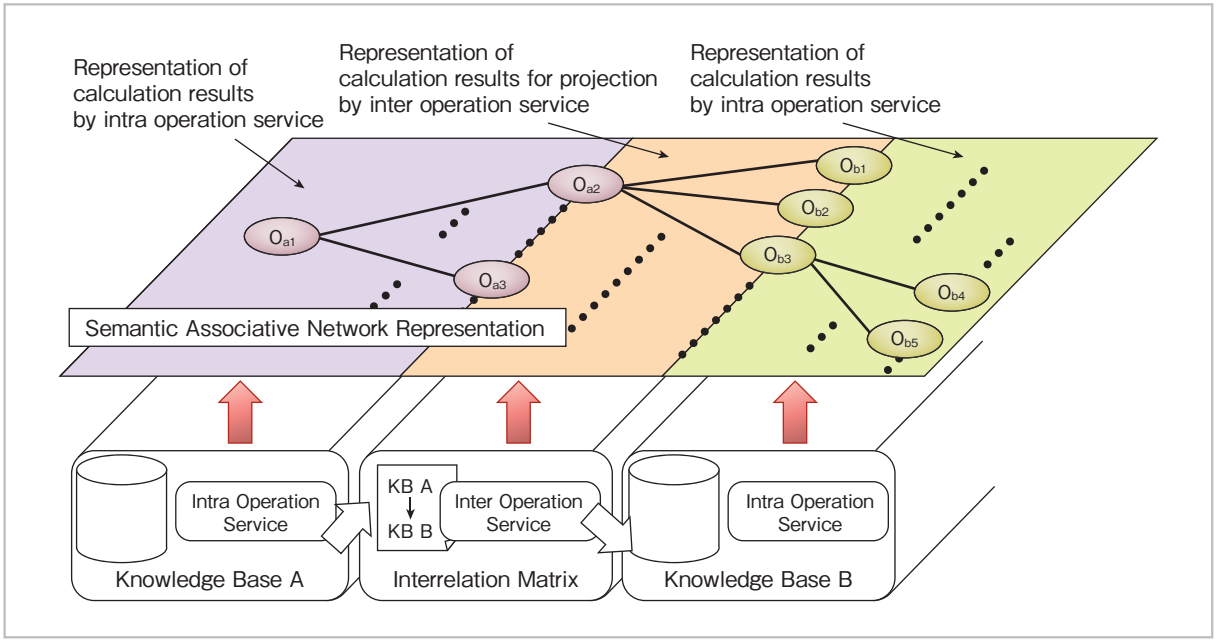

Fig.3 *An example of generating semantic correlation network by interconnecting heterogeneous knowledge bases through intra operation service and inter operation service*

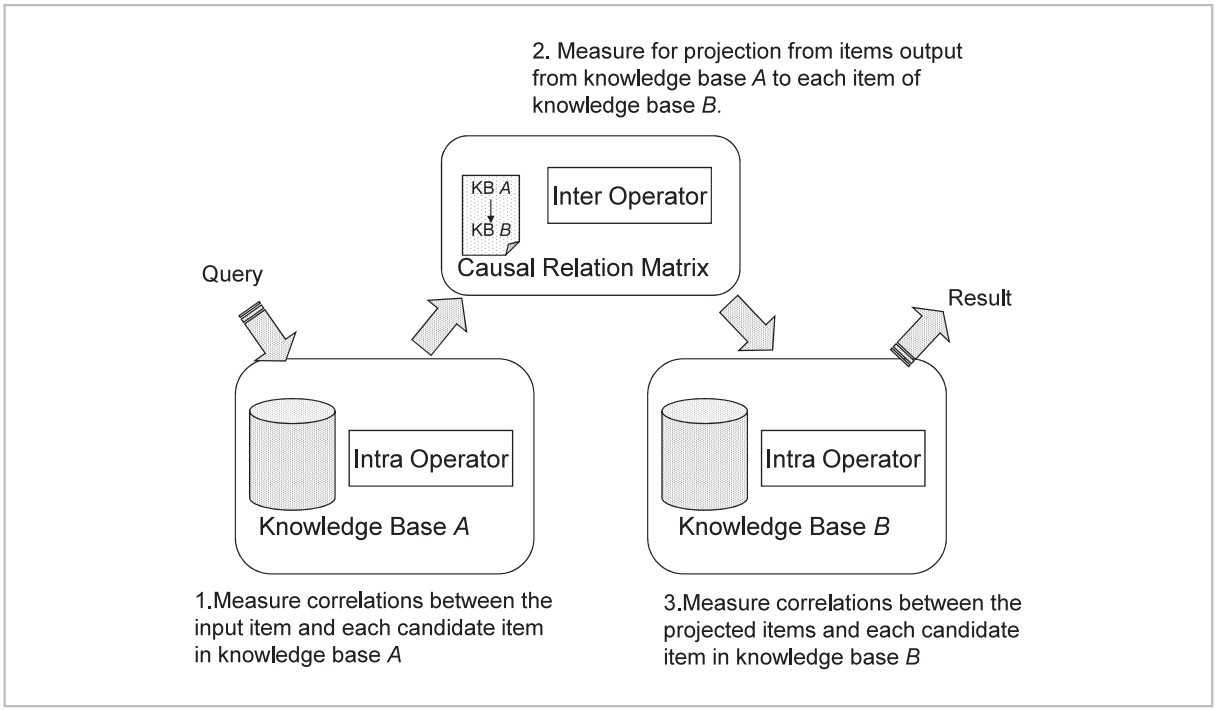

Fig.4 *Integration of heterogeneous knowledge bases through intra operation service and inter operation service*

Preparation: Generation of a correlation matrix

Inter operation service generates a correlation matrix used for a correlation search in advance. In the case of inter operation service, if there are m number of concept terms in knowledge base A and n number of concept terms in knowledge base B, the correlation matrix has a size of $m \times n$ matrix.

Step 1: Intra operation service for knowledge base A

Intra operation service in charge of knowledge service A does correlation research into knowledge base A. It acquires a set of concept terms that are highly correlated with a concept term given by a user.

Step 2: Inter operation service that maps knowledge base A to knowledge base B

Inter operation service does correlation search between knowledge base A and knowledge base B. Mapping a set of concept terms in knowledge base A acquired as search results in Step 1, through correlation matrix computation, to a set of concept terms in knowledge base B.

Step 3: Intra operation service for knowledge base B

It receives as input a set of concept terms in knowledge base B acquired from Step 2 and searches a set of concept terms in knowledge base B that are correlated with the input.

In summary, the system first expands the query with the use of concept terms contained in knowledge base A. Second, the system recalls the contents of knowledge base B that are highly correlated with these concept terms. Lastly, it determines the results with the use of concept terms contained in knowledge base B.

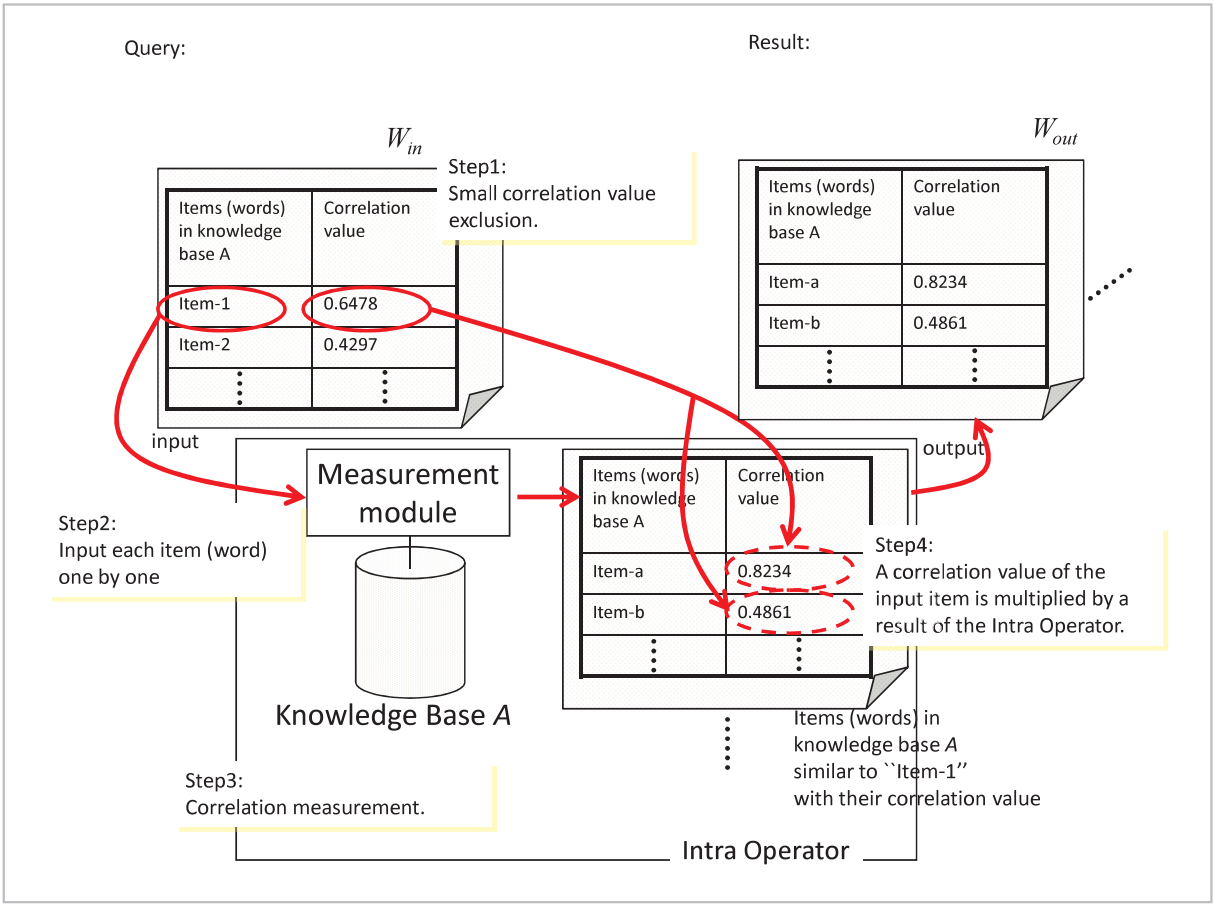

Fig.5 *Correlation search within knowledge base by intra operation service*

3.2 Details of Intra/Inter operation service

3.2.1 Intra operation service

Figure 5 illustrates the rough sketch of intra operation service. Correlation search by way of intra operation service is carried out in the following order.

Step 1: Concept terms given as the service input with low correlation coefficient should be considered as noise and removed accordingly.

Step 2: Send the input terms to a correlation measurement module individually.

Step 3: For each term, calculate the correlation coefficients with concept terms contained in knowledge base and select as candidates those concept terms that have higher correlation coefficients.

Step 4: Compute the correlation coefficients of search results, by multiplying the correlation coefficients of concept terms calculated in the Step 3.

The aforementioned framework is general and any correlation measurement method is applicable to the intra operation service. Our implementation employs a vector space model based on latent semantic indexing (LSI) [15] [16]. In this implementation, each individual concept term is represented by a feature vector which is constituted by a set of feature terms. In the end, knowledge base is represented by a data matrix [1] which combines them.

In order to create knowledge bases to be represented by a data matrix, we have developed a tool that makes use of Media Wiki. Media Wiki is a Wiki system famously used in Wikipedia (an online encyclopedia). Figure 6 shows the process of generating a data matrix using this tool.

Step 1: The contents created by Media Wiki are a Web page in which the heading and its explanation are written. The tool extracts the title and its explanation from each Media Wiki page and generates a dictionary in which the title becomes an entry word and the body of the text its explanation. The entry word corresponds to a concept term in knowledge base and becomes

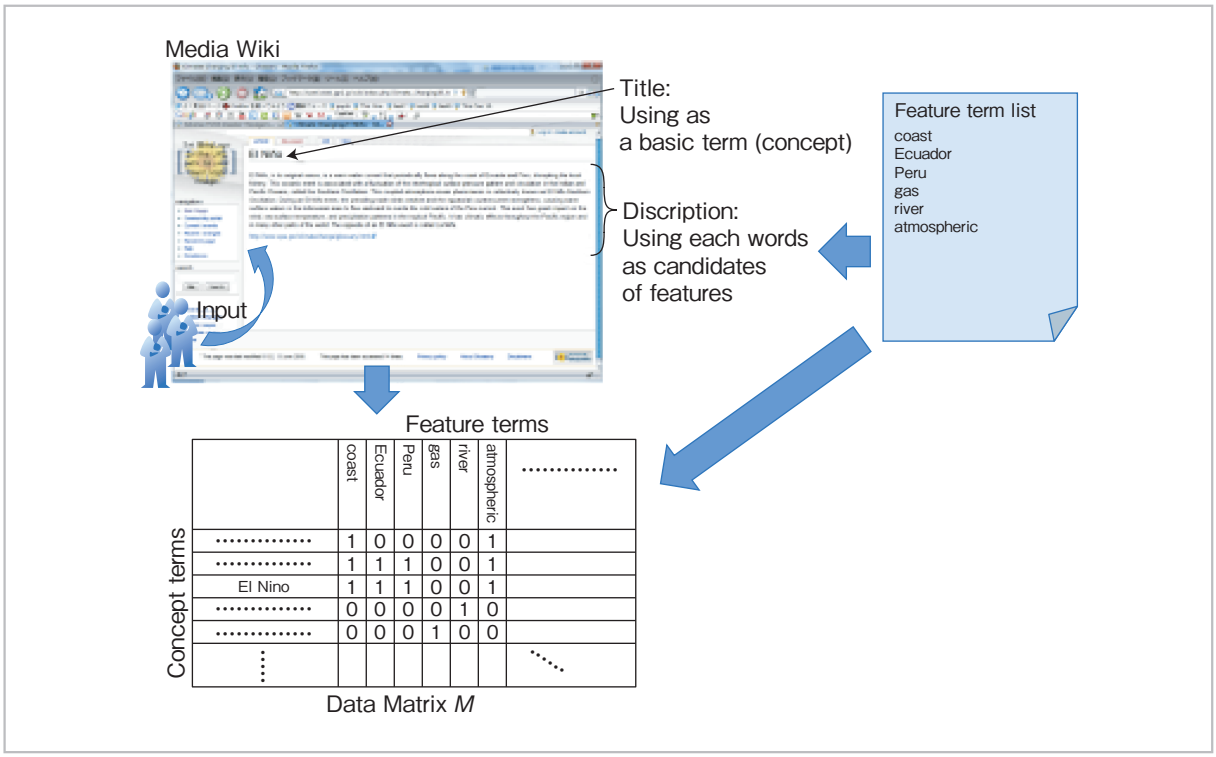

Fig.6 *Automatic generation of correlation matrix using MediaWiki system*

an individual node in the semantic correlation network.

Step 2: The tool prepares a list of feature terms that will become a base for characterizing concept terms. In our implementation, we use as feature terms a set of 2,000 base terms [1] extracted from an English dictionary.

Step 3: The tool characterizes a concept terms using the feature term vector. For each concept term created in Step 1, if its explanation contains a specific feature term, 1 is set to the corresponding vector element or otherwise it is set 0. In Figure 6, for a concept term (row) "El Nino" contained in data matrix M, set 1 to feature terms (column) such as "coast", "Ecuador" and "Peru" which appear in the body of its explanation and set 0 for other terms such as "gas" and "river". By combining these feature term vectors across a set of concept terms, the tool generates a data matrix. We have implemented a Media Wiki system which is embedded with a mechanism that automatically executes this process[24].

3.2.2 Inter operation service

Inter operation service does correlation search among heterogeneous knowledge bases. Figure 7 illustrates the execution order of inter operation service. Execution order is almost same as that of intra operation service. Only Step 3 is different as follows.

Step 3: Map a set of concept terms contained in search results from knowledge base A to a set of concept terms in knowledge base B. During the process, compute the new correlation coefficients as inner products of matrices along with the mapping.

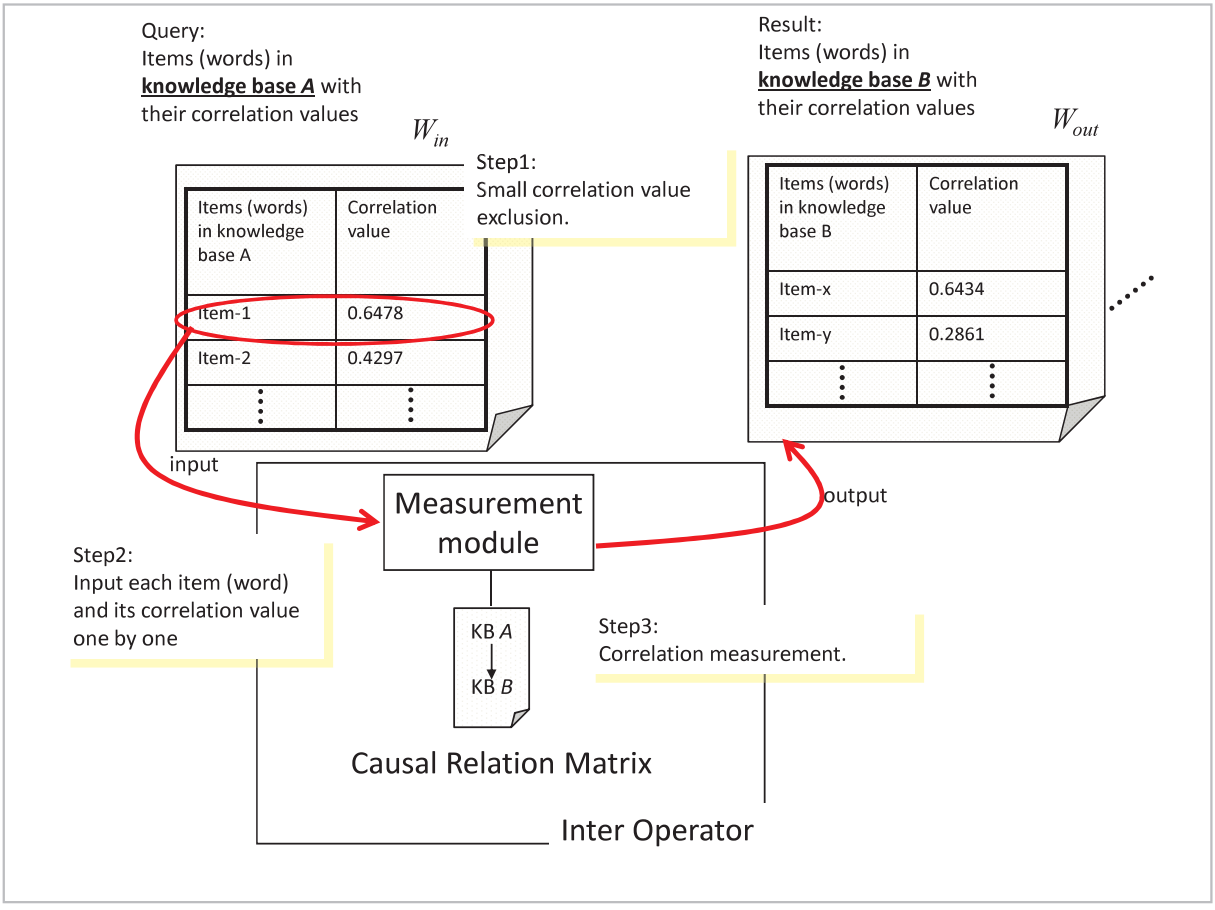

Fig.7 *Correlation search between heterogeneous knowledge bases by inter operation service*

For inter operation service to work, the generation of a correlation matrix is necessary. We have implemented a mechanism of automatically generating a correlation matrix from RSS data. RSS is broadly used to deliver Web news articles. Since RSS accommodates information from a variety of different domains, it is highly useful as an information source for generating a correlation matrix. Assembled RSS data are classified into groups based on topics. For instance, (see Fig. 8) in which RSS data are classified into 3 groups ("Unzen Fugendake", "Miyake Jima", "Sidoarjo Mudflow") and a correlation matrix is generated for each group. Each of these groups corresponds to a context in which inter operation service does correlation search (under a particular subject matter), which is used for combining heterogeneous knowledge bases under a variety of contexts and creating the semantic correlation networks.

The correlation matrix, which combines two knowledge bases (knowledge base A and knowledge base B), is generated in the following manner: First, select a set of concept terms from each of the knowledge bases. Next, for any combination of concept terms contained in each of these knowledge bases, reckon that there is a correlation between these concept terms and set 1 to the cell that corresponds to the combination in the correlation matrix, if the combination appears many times in the RSS data. Otherwise set to 0. The present operation is equivalent to searching from a vast amount of information on the Web for heterogeneous domains of evidences and discovering the axes for comparing them, which results in creation of a semantic space. That means if the inverse of inter operation service is computed, the reason why the link is established can be explained. This is the kind of result difficult to get with the traditional similarity searching and query expansion, and you can show from what the interconnection between heterogeneous domains is derived through this operation.

To mimic the relative comparison humans unconsciously make, our present method tries to find the axes of the relative comparison from the Web through inter operation and takes in similar evidences on the axis. Thus, our method is different from concept dictionary abstraction or a thesaurus (in natural language processing on this point) and it allows us to change the connection depending on different contexts.

At the present time, it looks as though our

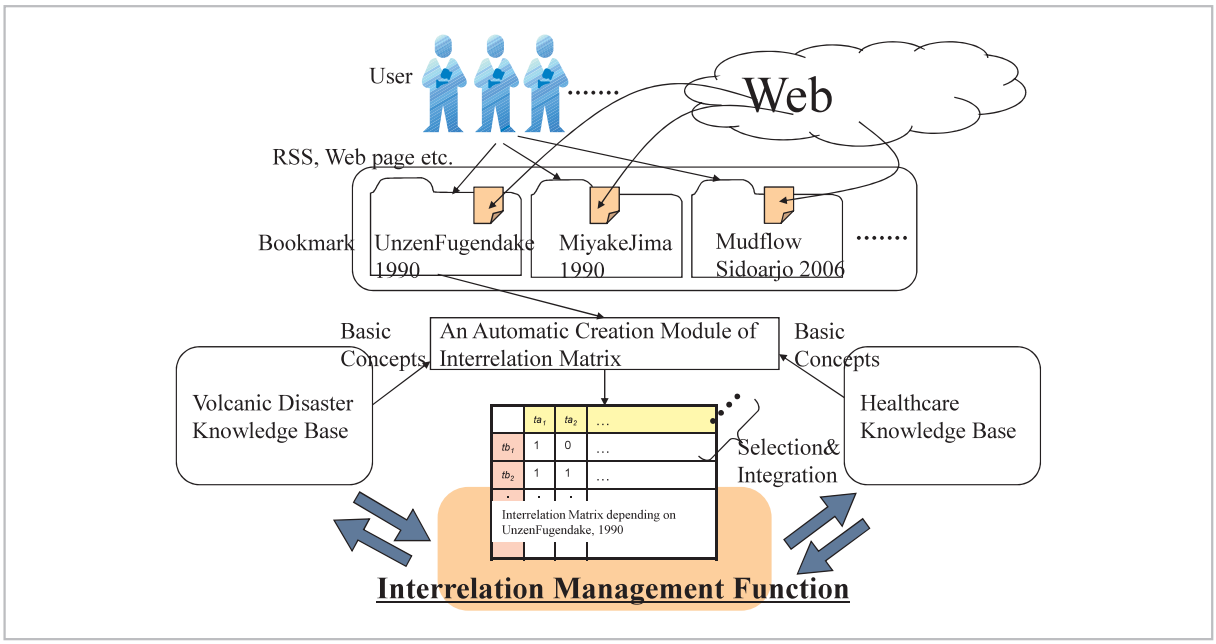

Fig.8 *Automatic generation of correlation matrix from RSS data*

architecture is a kind of query expansion which establishes (through intra operation as well as inter operation), the fact that, in response to a query which is a term *a* in knowledge base A, a term *b* in knowledge base B is combined with a term *m* in a correlation matrix M. But in the future, this architecture helps us cope with any query of the following kind. For example, our correlation search will find what kinds of domains are connected to the terms (words) that constitute the content of a certain news article. In general, from a term *m* in knowledge base M, it will find what kinds of knowledge bases are connected to it. As a model of human thought, the architecture presupposes that humans (users) have a solid guideline for comparison and look at all the different knowledge bases available.

Furthermore, the method we explained thus far is being realized in the word base. If these words are taken as a schema of the data base, it is possible for it to function as a schema mapping. And if our architecture realizes the correlation of a continuous value, we might be able to discover the connection not found in a concept dictionary or a thesaurus. This is a problem to solve in the future.

4 Experiment

4.1 Experiment environment

To assess the behavior of heterogeneous knowledge base integration (based on the proposed method), we have conducted experiments using real knowledge bases.

In our experiment, we have prepared knowledge bases of three different domains. First, we've created a volcanic disaster knowledge base from a book called "*Volcanic Hazards*" [20] and from pages on Wikipedia [21]. Second, we've created an environment knowledge from a list of terminologies on U.S. Environmental Protection Agency's homepage [19]. We also created a healthcare knowledge base derived from Wikipedia. In addition, we've created two correlation matrices, one connecting the volcanic disaster knowledge base to the environment knowledge base, the other connecting the volcanic disaster knowledge base to the healthcare knowledge base on the basis of information from relevant Websites [20]-[23].

We deploy LSI method as a correlation measurement method and generate a data matrix for each knowledge base according to the method mentioned in Subsection ***3.2.1***.

4.2 Experiment results

Figure 9 shows the results of generating a semantic correlation network with the use of a correlation matrix connecting the volcanic disaster knowledge base and the healthcare knowledge base in response to the query "volcanic gas". In this instance, starting from the left, there is a query "volcanic gas". From it, through the intra operation service (of the volcanic disaster knowledge base), concept terms such as "SO_2" in the volcanic disaster knowledge base are derived. Note that the number assigned to each edge stands for a correlation coefficient. Then, a correlation search from the volcanic disaster knowledge base to the healthcare knowledge base provides us with concept terms related to respiratory disorders such as "pulmonary edema", "bronchial asthma", "Yokkaichi asthma", and "obstructive lung disease", among others. Finally, the intra operation service of healthcare knowledge base provides concept terms that are highly correlated with these. As a result, we get concept terms in healthcare knowledge base that are highly correlated with the query "volcanic

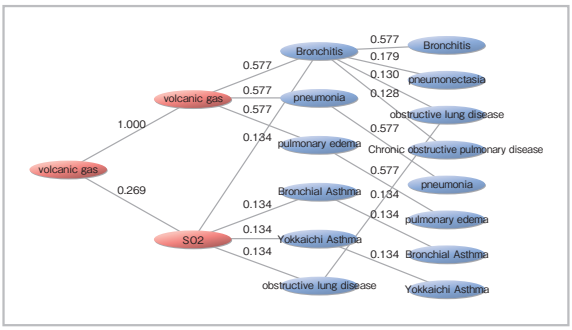

Fig.9 *Semantic correlation network representing correlation search results from volcanic disaster knowledge base to healthcare knowledge base*

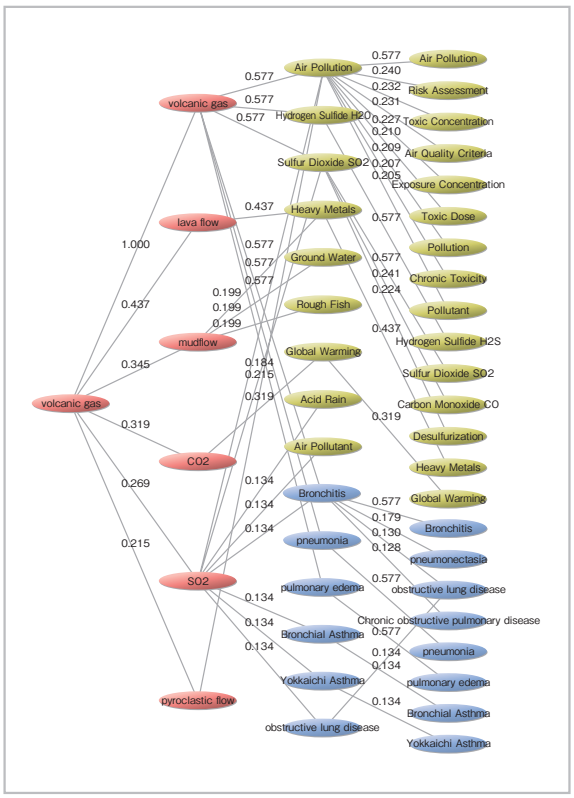

Fig.10 *Semantic correlation network representing correlation search results from volcanic disaster knowledge base to both environment knowledge base and healthcare knowledge base. (query: "volcanic gas")*

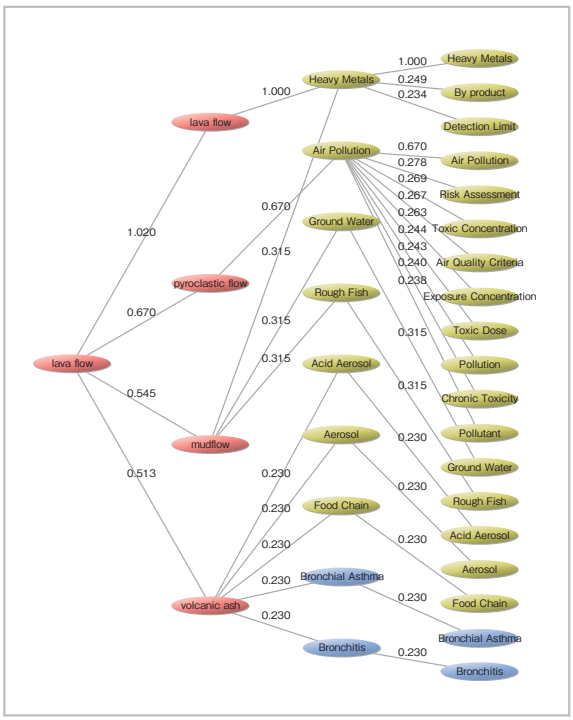

Fig.11 *Semantic correlation network representing correlation search results from volcanic disaster knowledge base to both environment knowledge base and healthcare knowledge base (query: "lava flow")*

gas" along with a semantic correlation network that shows the process of their derivations. In our framework, heterogeneous correlation search is made possible under a variety of different contexts by switching the correlation matrix.

Figure 10 shows the result of performing correlation searches into multiple domains with the use of multiple correlation matrices in response to the same query. It is the characteristic of our proposed method that by simply changing the combination of correlation matrix, it is able to perform correlation searches into many knowledge bases in parallel.

Figure 11 denotes a query that begins with "lava flow". Through the intra operation service of the volcanic disaster knowledge base, concept terms in volcanic disaster domain, such as "pyroclastic flow", "mudflow" and "volcanic ash" are derived. In the next step, these concept terms are expanded upon through inter operation services into highly correlated concept terms in the environment domain such as "heavy metals", "air pollution", "ground water", "rough fish", "acid aerosol", "aerosol" and "food chain", and in the health domain terms such as "bronchial asthma" and "bronchitis" are found. Lastly, through intra operation services that each correspond to an individual knowledge base, highly correlated concept terms in each domain are derived. As a result, we get concept terms in both environment and health domains that are highly correlated with the query "lava flow" along with semantic correlation networks that show their derivations.

5 Link-free browsing

We are developing link-free browsing system application that deploys the proposed method. Link-free browsing differs from tradi-

Fig.12 *Link-free browsing system*

tional Web browsing (through static hyperlink) because as a browsing method hyperlinks are generated dynamically based on a semantic correlation network (which represents correlation search results). In recent years, there has been an increase in the number of users who browse not only to obtain a Web page but to understand and learn about certain concepts. Traditional search engines, however, only search those pages that match the typed words and display a list of vast amounts of search results (without any arrangement). Under these conditions, it can be difficult for users who want to understand the concepts of a query. To make matters worse, for information on those knowledge domains on which a user is not an expert, the user may not even come up with appropriate terms and can miss important information. Our link-free browsing is expected to solve this sort of problems.

Figure 12 represents a rough sketch of link-free browsing system. Web browsing using a link-free browsing system is performed in the following manner:

1. **Content browsing mode:**
   The contents of a Web page are browsed as done by conventional Web browsers.
2. **View point select:**
   If a user selects and highlight an interested term from a Web page displayed in the content browsing mode, knowledge bases that are eligible for a correlation search (from the term) are prompted as "view points". If the user selects a view point, a corresponding correlation matrix is set to the system.
3. **Semantic correlation network browsing mode**
   The system performs correlation search into various domains from the given terms on the basis of view point and generates a semantic correlation network. Simultaneously, it thumbnails Web pages that correspond to individual concept terms, each of which is represented as a node in the semantic correlation network.

If the user clicks on a particular node, the system thumbnails the corresponding Web page. If you click on this thumbnail, a Web page is displayed and moves to content browsing mode.

Thus, link-free browsing (by switching between normal Web browsing) or the generation of dynamic semantic correlation network by way of correlation search, repeats Web page browsing. In addition, it searches for relevant concepts and helps deepen user's understanding of the target concept.

6 Summary

In this paper, we've explained the method of integrating heterogeneous knowledge bases developed on the Knowledge GRID. We've defined two sorts of Knowledge GRID services, intra operation service and inter operation service, necessary for the integration of heterogeneous knowledge bases. Moreover, we've proposed a method of generating a semantic correlation network using these services and of integrating knowledge bases. In addition, we've mentioned a link-free browsing system as an example of its application. With these two services, the traditional Web browsing, whose mainstream method was to search for Web pages that match the query keywords, can be changed into a method of obtaining relevant Web pages by discovering the correlation of the query with a variety domains of concept terms. We are aiming to restructure the traditional Web link structure based on the semantic correlation of contents.

This technology is a heterogeneous DB integration by correlation, which, unlike a concept dictionary or a thesaurus or ontology which abstracts from and infers from what is written, implements the human ability to think in context. The advent of cloud computing intimates that a variety of different clouds will be constructed by a variety of different institutes, organizations or communities who share similar thoughts and/or culture. There is the possibility of integrating these clouds and providing end users with new discoveries and better service on these occasions, correlation measurement will become increasingly important. This can be seen that the Knowledge GRID nodes in Fig. 2 will form a cloud.

Our future work focuses on the followings:

1) to shift the elements from data to services.
2) to propose the method of applying correlation measurement to service integration.
3) to realize distributed parallel processing of inter operation services.

Acknowledgements

We express our gratitude to Professor Yasushi Kiyoki, Faculty of Environment and Information Studies, Keio University, who has given us much important advice and direction in conducting this research.

References

1 T. Kitagawa and Y. Kiyoki, "The Mathematical Model of Meaning and its Application to Multidatabase Systems," Proc. 3rd IEEE International Workshop on Research Issues on Data Engineering: Interoperability in Multidatabase Systems, pp. 130–135, 1993.

2 Cannataro M. and Talia D., "The knowledge grid: Designing, building, and implementing an architecture for distributed knowledge discovery," Communications of the ACM 2003; 46(1): 89–93.

3 Zhuge H., "Communities and Emerging Semantics in Semantic Link Network: Discovery and Learning," IEEE Transactions on Knowledge and Data Engineering 2009; 21(6): 785–799.

4 Wilkinson R. and Smeaton A. F., "Automatic link generation," ACM Computing Surveys (CSUR) 1999; 31(4), No. 27.

5 Cleary C. and Bareiss R., "Practical methods for automatically generating typed links," Proceedings of the seventh ACM conference on Hypertext (HYPERTEXT '96), 31–41, 1996.

6 Stotts P. D. and Furuta R., “Dynamic adaptation of hypertext structure. Proceedings of the third annual ACM conference on Hypertext (HYPERTEXT ’91),” 219–231, 1991.
7 Armstrong R, Freitag D, Joachims T, and Mitchell T., “WebWatcher: A learning apprentice for the World Wide Web,” Proceedings of the 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous Distributed Environments 1995; AAAI Press.
8 Lieberman H., “Letizia: An Agent That Assists Web Browsing,” Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI ’95), 924–929, 1995.
9 Yan T. W, Jacobsen M, Garcia-Molina H, and Dayal U., “From user access patterns to dynamic hypertext linking,” Proceedings of the fifth international World Wide Web conference on Computer networks and ISDN systems 1996; 1007–1014.
10 Zettsu K, Nakanishi T, Iwazume M, Kidawara Y, and Kiyoki Y., “Knowledge cluster systems for knowledge sharing, analysis and delivery among remote sites,” Information Modelling and Knowledge Bases 2008; 19: 282–289.
11 Iwazume M, Kaneiwa K, Zettsu K, Nakanishi T, Kidawara Y, and Kiyoki Y., “KC3 Browser: Semantic Mashup and Link-free Browsing Proceedings of the 17th International World Wide Web Conference (WWW 2008),” 1209–1210, 2008.
12 Nakanishi T, Zettsu K, Kidawara Y, and Kiyoki Y., “Towards Interconnective Knowledge Sharing and Provision for Disaster Information Systems-Approaching to Sidoarjo Mudflow Disaster in Indonesia,”Proceedings of the 3rd Information and Communication Technology Seminar (ICTS2007), 332–339, 2007.
13 Miller R. J, Haas L. M, and Hernandez M. A., “Schema Mapping as Query Discovery,” Proceedings of the 26th International Conference on Very Large Data Bases (VLDB2000), 77–88, 2000.
14 Doan A. H, Madhavan J, Domingos P, and Halevy A., “Learning to Map between Ontologies on the Semantic Web,” Proceedings of the 11th international conference on World Wide Web, 662–673, 2002.
15 Berry M. W, Dumais S. T, and O’Brien G. W., “Using linear algebra for intelligent information retrieval,” SIAM Review 1995; 37(4): 573–595.
16 Deerwester S, Dumais S. T, Furnas G. W, Landauer T. K, and Harshman R., “Indexing by latent semantic analysis,” Journal of the American Society for Information Science 1990; 41(6): 391–407.
17 Ui T. Ed., “Volcanic Hazards,”University of Tokyo Press, 1997 (in Japanese).
18 Wikipedia (Japanese-language version), http://ja.wikipedia.org/
19 U.S. Environmental Protection Agency, http://www.epa.gov/
20 Global Volcanism Program, http://www.volcano.si.edu/
21 Volcano World, http://volcano.und.edu/
22 Environmental Protection, http://www.eponline.com/
23 MedicineNet.com, http://www.medicinenet.com/
24 Nakanishi, T., Zettsu, K., Kidawara, Y., and Kiyoki, Y., “SAVVY Wiki: A Context-oriented Collaborative Knowledge Management System,” Proc. of ACM Intl. Symp. on Wikis and Open Collaboration (Wikisym2009), P. 106, Oct. 2009.

(Accepted June 14, 2012)

***ZETTSU Koji**, Ph.D.*

Director, Information Services Platform Laboratory, Universal Communication Research Institute

Database, Data Engineering, Information Management, Information Retrieval

***NAKANISHI Takafumi**, Ph.D.*

Researcher, Information Services Platform Laboratory, Universal Communication Research Institute

Database, Multi Database System, Multimedia System, Intellectual Property Issues