A Semantic Search Space Integration Method for Meta-level Knowledge Acquisition from Heterogeneous Databases

Yasushi Kiyoki* and Saeko Ishihara**
*Faculty of Environmental Information
Keio University
**Graduate School of Media and Governance
Keio University
Fujisawa, Kanagawa 252-8520, Japan
phone: 81+466-47-5111, fax: 81-466-47-5041
e-mail: kiyoki@sfc.keio.ac.jp

Abstract.
In this paper, we present a semantic search space integration method for a heterogeneous database environment. This method realizes integration among various semantic search spaces for meta-level knowledge acquisition from heterogeneous databases. This semantic space integration is performed with the interpretation of meanings in terms of common concepts between different databases in heterogeneous research fields. In this paper, we also present an implementation method for applying our integration method to actual semantic search spaces. We have implemented an actual space integration system for accessing environmental and medical information resources. We clarify the feasibility and effectiveness of our method and system by showing several experimental results for environmental and medical document databases.

1 Introduction

The most important objective of our study is to develop a meta-level knowledge base system for realizing a creative environment in new research fields by integrating information resources in various research fields, such as cultural, social, and natural sciences. The meta-level knowledge base system leads to highly creative activities for human beings over various research fields by sharing, retrieving, editing and integrating databases through wide-area computer networks.

A number of legacy databases for individual scientific research fields are connected to wide-area computer networks. As the existing legacy databases for those research fields have been designed and created with their own data structures, data representations and languages, it is difficult to obtain global knowledge by sharing, retrieving, editing and integrating those databases. In such a heterogeneous database environment, semantic heterogeneity poses problems in integrating different databases. Knowledge sharing, semantic retrieving and integrating those databases are essentially important for dynamically creating new research fields over various research fields[4, 9, 10, 13]. We have proposed a meta-level database system which realizes an intelligent database integration environment[4, 6, 8]. In this system, databases for various fields are connected to the meta-level layer, and those databases are integrated by semantic functions. By the connection among those databases from different fields, this system provides a knowledge integration environment for new scientific research related to various scientific fields.

We have also proposed a metadatabase system with a new semantic associative search method based on a mathematical model of meaning (MMM) [2, 6]. This method

makes it possible to extract and obtain significant information from multidatabases with a machinery for semantic associative search. In this method, the acquisition of information in multidatabases is performed by semantic computations.

It is complicated to deal with the meanings of data items in a multidatabase environment. One of the hardest problems is that it is difficult to identify the semantic equivalence, similarity and difference between data items which are extracted from different databases [1, 6, 9, 14]. It is not easy for users to select the appropriate databases and extract significant information for their requests. To provide the facilities for selecting the appropriate databases and extracting the significant information from those databases, a methodology for realizing semantic interoperability is an important part of database integration technology [1, 2, 10, 14]. The problematic relationships between data items for realizing semantic interoperability are classified into two types: "homonyms" and "synonyms." Homonym means the same data item is used for different concepts. Synonym means the same concept is described by different data items in different databases.

Our mathematical model of meaning (MMM) provides a function to compute the semantic equivalence, similarity and difference between data items which are included in different databases and realizes semantic interoperability among the data items. This model is used to find semantically equivalent or similar data items with different data representations and to recognize the different meanings of a data item. The main feature of this model is that the specific meaning of a data item can be dynamically fixed and unambiguously recognized according to the context. In this method, the data items of multidatabases are mapped into an orthogonal image space and selected by an intelligent semantic associative search mechanism[2, 6].

Several information retrieval methods, which use the orthogonal space created by mathematical procedures like SVD (Singular Value Decomposition), have been proposed. The MMM is essentially different from those methods using the SVD (e.g. the Latent Semantic Indexing (LSI) method [15, 16]). The essential difference is that the MMM provides the important function for semantic projections which realizes the dynamic recognition of the context. That is, in our model, the context-dependent interpretation is dynamically performed for computing the distance between words by selecting a subspace from the entire orthogonal semantic space. In our model, the number of phases of the contexts is almost infinite (currently 2^{2000} and 2^{800}, approximately). Other methods do not provide the context dependent interpretation for computing equivalence and similarity in the orthogonal space, that is, the phase of meaning is fixed and static.

We have applied the MMM to several multimedia database applications, such as image and music data retrieval by impressionistic classification. We have introduced these research results in [2, 6] and the book "Multimedia Data Management - using metadata to integrate and apply digital media -," McGraw Hill, Chapter 7, 1998 [5]. Through these studies, we aim to create a new meta-level knowledge base environment by applying those methods to data retrieval, data integration and data mining [3, 7].

In this paper, we present a new method of semantic retrieval space integration(SSI) for heterogeneous fields. This method makes it possible to integrate semantic retrieval spaces with the interpretation of meanings by using common concepts (common terms) for matrices of heterogeneous fields. This method realizes the information retrieval from viewpoints related to semantically integrated fields. In this paper, we also present an implementation method for applying our integration method to semantic associative search spaces. We clarify the feasibility and applicability of our method by several experiments for environmental fields.

In this method, it is assumed that common concepts (common terms) between heterogeneous fields are detected in advance before applying this method to the semantic associative search spaces corresponding to those fields. It is assumed that the semantic equivalence and similarity between terms in different fields are recognized by using our mathematical model of meaning (MMM) or the concept of ontology[1, 6, 9, 14].

The SSI method is used for integrating orthogonal spaces created by mathematical procedures like MMM[2, 6] and SVD(Singular Value Decomposition: e.g. the Latent Semantic Indexing (LSI) method [15, 16]). It can be applied to various semantic information retrieval methods using vector spaces. In those methods, a vector space is created for information retrieval for a single field or several fields which are fixed in advance. Although it is possible to create several vector spaces for several fields, those vector spaces are not integrated into a single space. Our SSI method dynamically integrates arbitrary vector spaces into a single space for realizing knowledge acquisition from information resources related to various research fields.

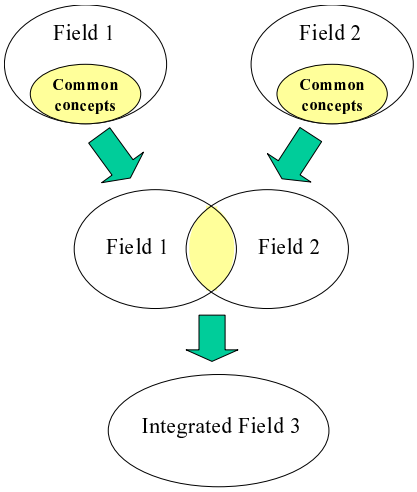

Fig.1: Integrated fields by using common concepts

2 Outline of the Semantic Space Integration Method

We propose a new semantic space integration method (SSI) for obtaining information related to multiple research fields. This method realizes semantic search space integration from different semantic search spaces as shown in Fig. 1. The procedure for semantic space integration and semantic search consists of the following processes:

Process-1: Creation for individual matrices (Original matrix creation),

Process-2: Semantic Space Integration (SSI) for matrices (Space Integration),

Process-3: Orthogonal semantic space creation for MMM and SVD (Integrated and Orthogonal Semantic Space creation).

Our SSI method defines a set of functions and data structures to realize Process-2.

Process-1 and Process-3 are dependent on semantic search methods. In Section 4 we explain Process-1 and Process-3 in the case for applying the mathematical model of meaning (MMM) to orthogonal semantic space creation.

2.1 Data structure

The data structure is defined as a set of basic words and features in the form of a matrix with basic words and feature words as shown in Fig. 2. The data structure is referred to as "space matrix with words".

A set of basic words which characterizes the data items to be used is given in the form of an m by n matrix. That is, for given m basic words, each word is characterized by n features.

2.2 Basic function for semantic space integration

The semantic space integration function is defined for integrating individual spaces corresponding to two different fields. This method can be applied to integration of various semantic spaces created in different research fields independently of the order of the space sequence in semantics.

This function integrates two space matrices as shown in Fig.3 and Fig.4. That is, this function performs the semantic space integration between two different research fields. This new function is very important for integrating semantic spaces originally created in different research fields independently.

Although this function is not commutative between two space matrices M1 and M2, the integrated space M3 is not dependent on the order of the space sequence in semantics. That is, the order of the result spaces does not change the semantics in the

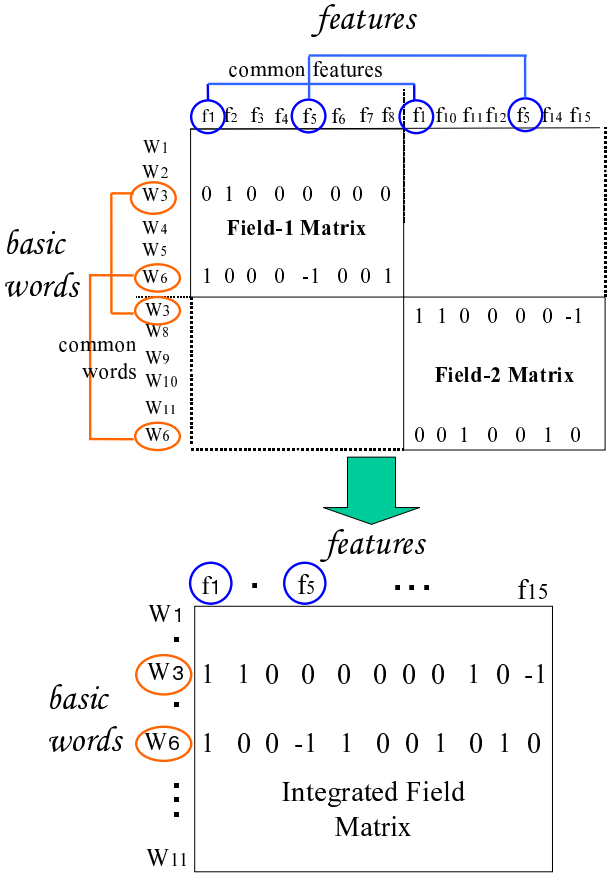

Fig.2: Integrated matrices

integrated space. Therefore, this function can be applied repeatedly to integration of various semantic spaces.

This function consists of the following three steps:

Step-1: Feature word integration:

Each feature word in the space matrix M2 is checked whether it exists commonly in the feature words of the space matrix M1 with the interpretation of synonymy. It is assumed that the semantic equivalence and similarity between words are recognized in advance by using ontology research results, such as in [1, 6, 9, 14], before applying this step to the space matrices M1 and M2. If a synonym or a common concept exists between feature words in M1 and M2, it is removed from the set of feature words of M2. The feature words of the integrated space matrix M3 consist of the feature words of M1 and the reduced feature words of M2, as shown in Fig. 3 and 4.

Step-2: Basic word integration:

Each basic word in the space matrix M2 is checked whether it exists commonly in the basic words of the space matrix M1 with the interpretation of synonymy. If a synonym or a common concept exists between basic words in M1 and M2, it is removed from the set of basic words of M2. The basic words of the integrated space matrix M3 consists of the basic words of M1 and the reduced basic words of M2, as shown in Fig. 3 and 4.

Step-3: Value settings to the integrated space matrix M3:

The basic words and feature words are set as vertical and horizontal words in M3 as shown in Fig. 4. Each element of M3 is set in this step. M3 consists of four submatrices M1', M2', M2" and M2"' as shown in Fig. 4. M3 is the matrix integrating two different space matrices which are created independently from different research fields.

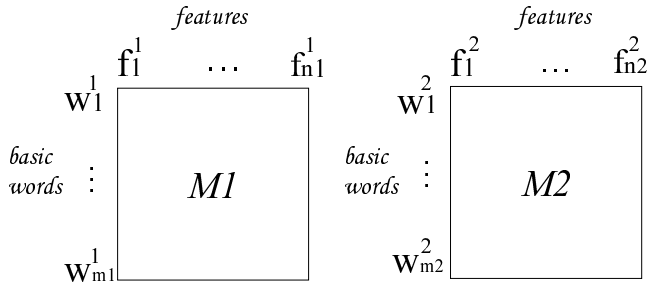

Fig.3: Two Original Single-Field Matrices (M1, M2)

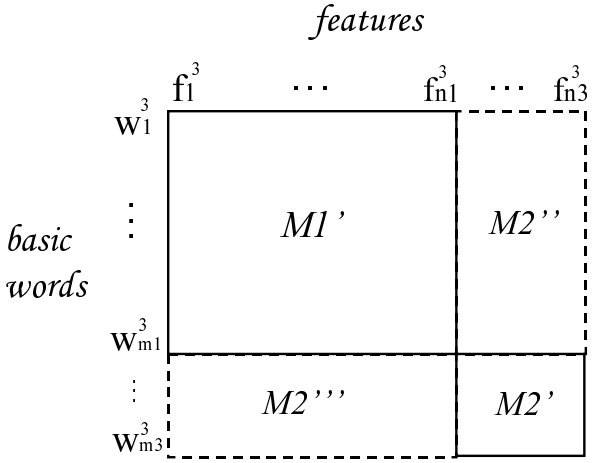

Fig.4: Integrated Matrix M3

M1' is the submatrix corresponding to the original M1. The basic and feature words in M1' are the same as those words in M1. Each element of M1' is set to the same value as the value in the original M1 if the basic word and the feature word corresponding to this element are not commonly existing between the original M1 and M2. If both of the basic word and the feature word in M1' are commonly existing between the original M1 and M2, the element corresponding to these words is set to the value computed by the integrating operation(integrator)between the M1 and M2 elements corresponding to the common basic and feature words.

M2' is the submatrix corresponding to the reduced M2 after eliminating both of common basic and feature words to M1 from M2. Each element of M2' is set to the same value as the value in the original M2 in terms of the reduced basic and feature words neither of which has common words to the original M1.

M2" is the submatrix corresponding to the elements where the common basic words are existing between the original M1 and M2 and the feature words between M1 and M2 are different. Each element of M2" is set to the same value as the value corresponding to the basic and feature words in the original M2.

M2"' is the submatrix corresponding to the elements where the common feature words are existing between the original M1 and M2 and the basic words between M1 and M2 are different. Each element of M2"' is set to the same value as the value corresponding to the basic and feature words in the original M2.

3 Integration Function

In this section, we define the integration function for our semantic space integration model. The data structures and the basic operations explained in Section 2 are formulated in the following expressions.

The semantic space integration function F is expressed as

$$M3 = F(M1, M2)$$
$$\{e^3_{i,j}\} = F(\{e^1_{i,j}\}, \{e^2_{i,j}\}),$$

where $e^1_{i,j}$ and $e^2_{i,j}$ are elements of the original space matrices M1 and M2, respectively, and $e^3_{i,j}$ is an element of the integrated space matrix M3.

The space matrix integration function F is expressed in the following formulation.

The schema of the space matrix M3 is represented as an ordered set of basic words and an ordered set of feature words as shown in Fig. 4. The ordered set of basic words is the vertical elements in M3, and the ordered set of feature words is the horizontal elements in M3.

The feature words in the space matrix M3 to be created as an integrated space matrix are extracted from the feature words of the original space matrices M1 and M2. $\bar{O}Set$ is the operator which makes an ordered set for words. The numbers of feature words in M1 and M2 are represented as $n1$ and $n2$, respectively. The number of feature words in M3 is represented as $n3$,

$$n2' = count\ (Difference\ (Set_{i=1,n2}(f^2_i), Set_{j=1,n1}(f^1_j))$$
$$\bar{O}Set_{k=1,n2'}(f^{2'}_k) \equiv$$
$$\quad Difference\ (\bar{O}Set_{i=1,n2}(f^2_i), \bar{O}Set_{j=1,n1}(f^1_j))$$
$$n3 = n1 + n2'.$$

The basic words in the space matrix M3 are extracted from the basic words of the original space matrices M1 and M2. The numbers of basic words in M1 and M2 are represented as $m1$ and $m2$, respectively. The number of basic words in M3 is represented as $m3$,

$$m2' = count\ (Difference\ (Set_{i=1,m2}(w^2_i), Set_{j=1,m1}(w^1_j))$$
$$\bar{O}Set_{k=1,m2'}(w^{2'}_k) \equiv$$
$$\quad Difference\ (\bar{O}Set_{i=1,m2}(w^2_i), \bar{O}Set_{j=1,m1}(w^1_j))$$
$$m3 = m1 + m2'.$$

The integrated matrix M3 consists of four submatrices M1', M2', M2" and M2"'.

The matrix M1' is the submatrix corresponding to the original M1. The basic and feature words in M1' are the same as those words in M1. Each element $e^{1'}_{i,j}$ of M1' is defined as follows:

$$e^{1'}_{i,j} = \begin{cases} integrator((e^1_{i,j}), (e^2_{i',j'})) & if((i \le n1) \wedge (j \le m1)) \wedge (f^1_i = f^2_{i'}) \wedge (w^1_j = w^2_{j'}) \\ e^1_{i,j} \qquad \text{otherwise} \end{cases}$$

The domain of the matrix elements and the integrator are application-dependently fixed. This model does not give restriction for defining them. In our current application study to environmental and medical semantic spaces, the domain of the elements is {-1, 0, 1}, and the integrator is defined as a three-valued logical OR operator where the OR operator between "-1" and "1" gives "0", and that between "-1" and "0" gives "-1".

M2" is the submatrix corresponding to the elements where the common basic words are existing between the original M1 and M2 and the feature words between M1 and M2 are different. Each element $e^{2''}_{i,j}$ of M2" is defined as follows:

$$e^{2''}_{i,j} = \begin{cases} e^2_{i',j'} & if((i > n1) \wedge (j \le m1)) \wedge (w^1_{j'} = w^2_{j'}) \\ \emptyset \qquad \text{otherwise} \end{cases}$$

M2"' is the submatrix corresponding to the elements where the common feature words are existing between the original M1 and M2 and the basic words between M1 and M2 are different. Each element $e^{2'''}_{i,j}$ of M2"' is defined as follows:

$$e^{2'''}_{i,j} = \begin{cases} e^{2}_{i'j'} & if((i \leq n1) \wedge (j > m1)) \wedge (f^1_{i'} = f^2_{i'}) \\ \emptyset & \text{otherwise} \end{cases}$$

M2' is the submatrix corresponding to the reduced M2 after eliminating both of common basic and feature words to M1 from M2. Each element $e^{2'}_{i,j}$ of M2' is defined as follows:

$$e^{2'}_{i,j} = \begin{cases} e^{2}_{i'j'} & if((i > n1) \wedge (j > m1)) \wedge (f^1_i \neq f^2_{i'}) \wedge (w^1_j \neq w^2_{j'}) \end{cases}$$

The space matrix M3 is created by combining the submatrices M1', M2', M2" and M2"'. Each element $e^3_{i,j}$ of M3 is defined as follows:

$$e^3_{i,j} = \begin{cases} e^{1'}_{i,j} & if(i \leq n1) \wedge (j \leq m1) \\ e^{2}_{i,j} & if(i > n1) \wedge (j \leq m1) \\ e^{2}_{i,j} & if(i \leq n1) \wedge (j > m1) \\ e^{2'}_{i,j} & if(i > n1) \wedge (j > m1) \end{cases}$$

4 Outline of the Mathematical Model of Meaning

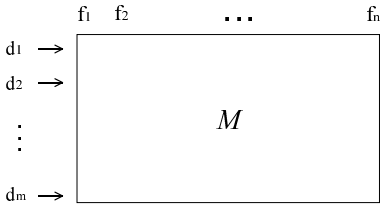

Fig.5: Metadata represented in data matrix M

In this section, the outline of the mathematical model of meaning (MMM) is briefly reviewed. This model has been presented in [2, 6] in detail.

In MMM, Process-1 (Creation for an individual matrix) and Process-3 (Orthogonal semantic space creation) are realized by the following methodology. Process-2 (Semantic Space Integration(SSI) for matrices) is realized by the function defined in Section 2.

1. Assumption :

   As Process-1 in Section 2, a set of basic words which characterizes the data items to be used is given in the form of an m by n matrix. That is, for given m words, each word is characterized by n features, as shown in Fig. 5.

2. Defining the image space $\mathcal{I}$:

   As Process-3 in Section 2, first we construct the correlation matrix with respect to the features. Then we execute the eigenvalue decomposition of the correlation matrix and normalize the eigenvectors. We define the image space $\mathcal{I}$ as the span of the eigenvectors which correspond to nonzero eigenvalues. We call such eigenvectors semantic elements hereafter. We note that since the correlation matrix is symmetric, the semantic elements form orthonormal bases for $\mathcal{I}$. The dimension ν of the image space $\mathcal{I}$ is identical to the rank of the data matrix A. Since $\mathcal{I}$ is ν dimensional Eucledian space, various norms can be defined and a metric is naturally introduced.

3. Defining a set of the semantic projections Π_ν:

   We consider the set of all the projections from the image space $\mathcal{I}$ to the invariant subspaces (eigen spaces). We refer to the projection as the semantic projection

and the corresponding projected space as the semantic subspace. Since the number of i dimensional invariant subspaces is $(\nu(\nu-1)\cdots(\nu-i+1))/i!$, the total number of the semantic projections is 2^ν. That is, this model can express 2^ν different phases of the meaning.

4. Constructing the Semantic Operator S_p:

   Suppose a sequence s_ℓ of ℓ words (context words) which determines the context is given. We construct an operator S_p to determine the semantic projection according to the context. Context words are given as a sequence of several keywords which are defined with n-dimensional vectors to specify the query for information retrieval. Several examples for context words and vectors were presented in [4, 6].

   We call the operator a semantic operator.

   (a) First we map the ℓ context words in databases to the image space $\mathcal{I}$. This mathematically means that we execute the Fourier expansion of the sequence s_ℓ in $\mathcal{I}$ and seek the Fourier coefficients of the words with respect to the semantic elements. This corresponds to seeking the correlation between each context word of s_ℓ and each semantic element.

   (b) Then we sum up the values of the Fourier coefficients for each semantic element. This corresponds to finding the correlation between the sequence s_ℓ and each semantic element. Since we have ν semantic elements, we can constitute a ν dimensional vector. We call the vector normalized in the infinity norm the semantic center of the sequence s_ℓ.

   (c) If the sum obtained in (b) for a semantic element is greater than a given threshold ε, we employ the semantic element to form the projected semantic subspace. We define the semantic projection by the sum of such projections.

   This operator automatically selects the semantic subspace which is highly correlated with the sequence s_ℓ of the ℓ context words which determines the context.

   This model makes dynamic semantic interpretation possible. We emphasize here that, in our model, the "meaning" is the selection of the semantic subspace, namely, the selection of the semantic projection and the "interpretation" is the best approximation in the selected subspace.

As an example, we have implemented an experimental system of the mathematical model of meaning [4, 6, 8]. As an example of the m by n matrix, we used basic words in an English dictionary "General Basic English Dictionary [12]" in which approximately 871 basic words are used to explain each English word[12]. Those English words are used as features, that is, they are used as the features corresponding to the columns in the matrix. That is, 871 features are provided to make the image space. And, approximately 2115 words are used to represent the words corresponding to the rows in the matrix. Those words are used as the basic words in the English dictionary "Longman Dictionary of Contemporary English [11]." The 2115×871 matrix is used to create the image space. By using this matrix, an image space is computed within the framework of the mathematical model of meaning. This space represents the semantic space for computing meanings of the keywords and data items which are used in a multidatabase environment. A given keyword and data items are mapped into this space, and semantic equivalence and similarity between the keyword and data items are computed in a mathematical way. This image space consists of 868 dimensional orthogonal vectors, approximately.

Similarly, the 2115×2115 matrix is used to create another image space. In this matrix, approximately 2115 words of "Longman Dictionary of Contemporary English [11]" are used to represent not only the basic words corresponding to the rows but also the features the columns in the matrix. This image space consists of 2000 dimensional orthogonal vectors, approximately. In those image spaces, the number of phases of the contexts is almost infinite (currently 2^{868} and 2^{2000}, approximately).

5 Application to Integration for two Semantic Spaces for Environmental Information and Medical Information

To clarify the feasibility and effectiveness of our space integration method, we performed several experiments by using two different semantic spaces which include environmental

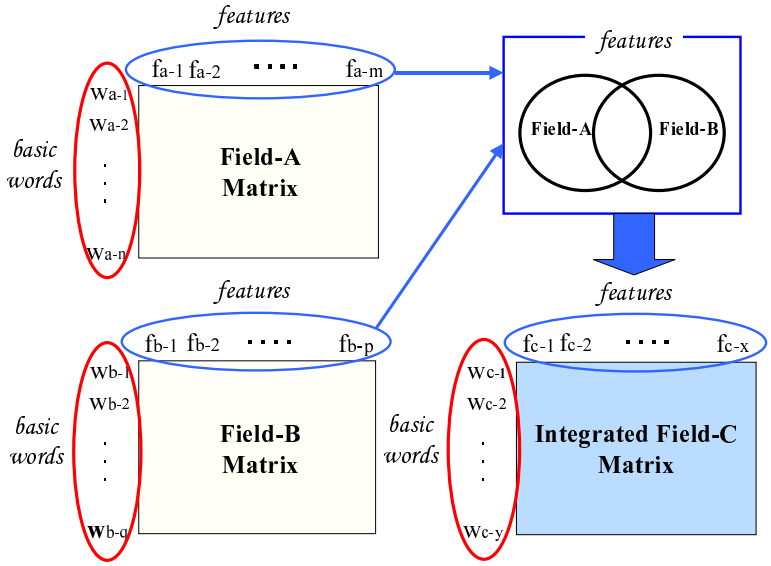

Fig.6: Integration for features and words

and medical information, respectively. The objective of these experiments is to evaluate the effectiveness and applicability of our method to combine the actual different semantic spaces. We have actually created two semantic spaces for environmental and medical information, and integrated those spaces by applying our space integration method.

5.1 Experimental environment

As a multidatabase environment including environmental and medial information, the semantic spaces for those fields were created and integrated into an integrated semantic space for an environmental and medical information by using our SSI method, as shown in Fig. 6 and 7.

For creating the environmental semantic space, we have referenced a dictionary for the environmental field [17, 18, 19]. From the dictionary, we have extracted basic words and feature words for the environmental semantic space(environmental metadata space).

For the medial semantic space, we have referenced a dictionary for the medical field [20, 21] and extracted basic words and feature words for the medical semantic space(medical metadata space).

Three semantic spaces(M1, M2 and M3) have been created for retrieving documents related to environmental, medical and integrated environmental-medical semantic spaces. Each semantic space is independently used for retrieving those documents.

(1) environmental semantic space for retrieving environmental documents as the space matrix M1 in Section 2.

(2) medical semantic space for retrieving medical documents as the space matrix M2 in Section 2.

(3) integrated environmental-medical semantic space for retrieving environmental and/or medical documents as the space matrix M3 in Section 2.

Our space integration method is applied to create the integrated environmental-medical semantic space (M3) from the space matrices M1 and M2. We have implemented our space integration method by developing the programs in Perl programming language.

The space matrices for those spaces are shown in Table 1.

As retrieval candidate documents for the environmental semantic space, we mapped 500 newspaper articles into the space.

As retrieval candidate documents for the medical semantic space, we mapped 500 newspaper articles into the space.

We also mapped those 1000 documents into the integrated environmental-medical semantic space.

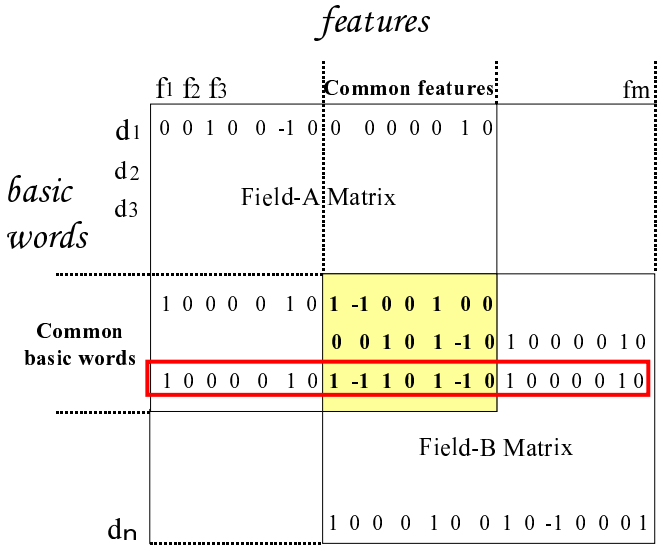

Fig.7: Integration for vector elements

Table 1: Semantic spaces

	Number of feature	Number of words	Space dimension
Environmental semantic space	425	469	415
Medical semantic space	437	690	436
Integrated semantic space	806	1127	794
Number of common terms	56	32	-

For mapping each of those documents into semantic spaces, we created a vector for each document by extracting a set of metadata from the document. Several metadata sets are shown in Table 2 as examples.

5.2 Experiment-1

In Experiment-1, we evaluate the quality of the retrieval results in a single semantic space created from a single field. That is, the environmental semantic space mapping the 500 environmental documents is used for evaluating retrieval results in this space. Similarly, the medical semantic space is used for evaluation.

5.2.1 Evaluation method

In Experiment-1, 15 queries have been given to our experimental system with the environmental semantic space. Similarly, other 15 queries have been given to our experi-

Table 2: Examples of metadata

Document-ID	metadata
980122259	influenza, virus, infection, tuberculosis, cold
980210202	cancer, virus, gene, inflammation, infection, hepatitis, hepatic, - -
971023101	dioxin, hormone, environment, pesticide
980623148	ulcer, cancer, gastritis, antibiotic, bowel
980723299	stress, hormone, excitation, diabetes, brain, obesity, hospital, senescence
001224107	greenhouse gas, warming, sea, limate change, Kyoto protocol, Forest, - -
001117222	dioxin, contamination, environmental quality, heavy metal, soil pollution
000528158	forest, bionomics, ecosystem, tree planting, forestry
980422027	ozone, ozone layer, chlorofluocarbon, greenhouse effect, warming, - -
981207252	contamination, sewer, river, environmental hormones, lake, water pollution, - -

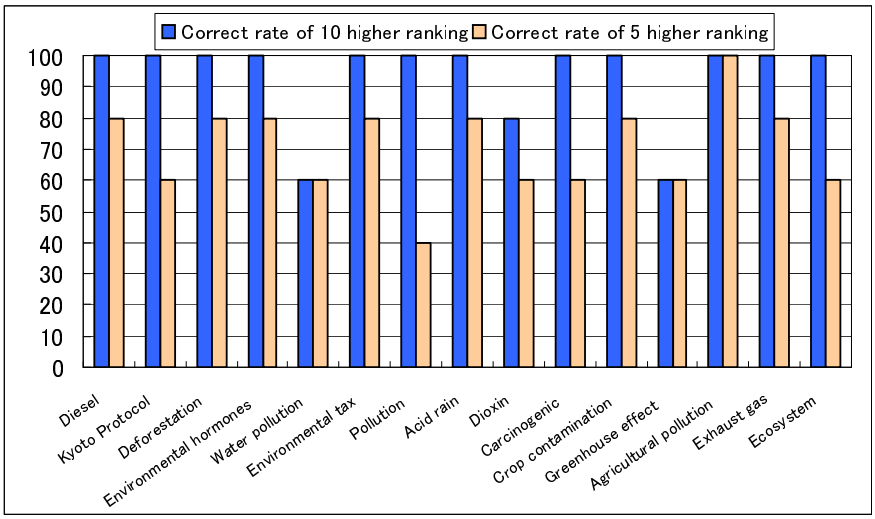

Fig.8: The result for the Environmental Space

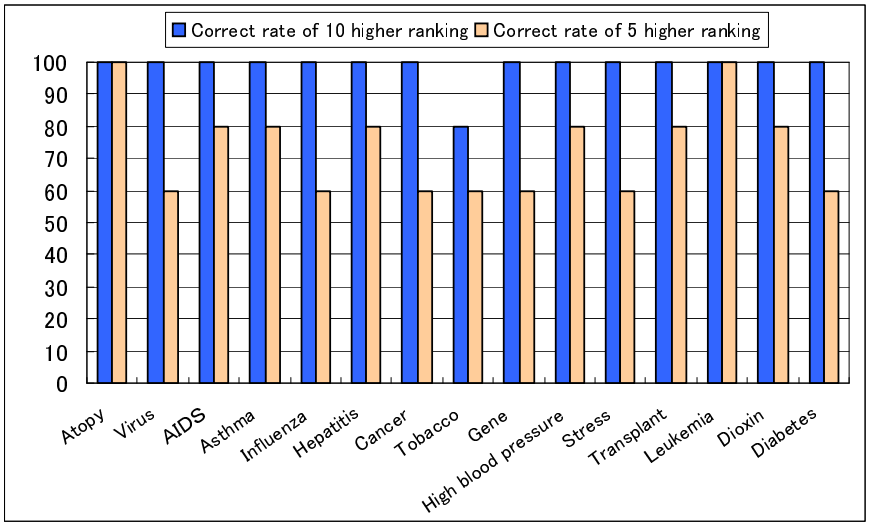

Fig.9: The result for the Medical Space

mental system with the medical semantic space. We have fixed 5 documents as correct answers for each query in advance. In the experimental results, we show the ratio (recall evaluator) of the correct answers included in the top 5 and top 10 retrieval results in Fig.8 for the environmental semantic space, and in Fig.9 for the medical semantic space.

5.2.2 Experimental results

Those experimental results show that our system realizes high quality retrieval for documents in terms of the recall evaluator. For all queries, the ratio (recall ratio) of the correct answers included in the top 10 retrieval results is in 80 to 100%. And, for a half of queries, the ratio (recall ratio) of the correct answers included in the top 5 retrieval results is in more than 80%.

In the query "environmental hormones", the document, whose metadata do not include "environmental hormones" itself, is selected in the high ranking in the retrieval results. The metadata of this document are "PCB ,hormone, contamination, agricultural chemicals, insecticide, industrial waste". This experiment shows that our semantic retrieval system realizes the dynamic computation for semantic correlation between a

given query word and metadata of documents.

In Experiment-2, we evaluate the quality of retrieval results in the integrated semantic space created by our semantic space integration method, in comparing with retrieval results in the single semantic spaces. The main advantage of our method is that it realizes the acquisition of the documents which are related to two or more different fields, but are not selected in each single semantic space because they do not have high correlation to each single field.

We mapped 1000 documents into the integrated environmental-medical semantic space. Those documents, which are used in Experiment-1, are related to the environmental and/or medical fields.

5.2.3 Evaluation method

In this experiment, the same queries are given to the environmental semantic space, the medical semantic space and the integrated environmental-medical semantic space, respectively.

The experimental results are shown in Tables 3, 4 , and 5.

The left-side numbers in each column are document identifiers (Document ID's), and the right-side values represent semantic correlations between a given query and each document.

5.2.4 Experimental results

Table 3 shows the retrieval results to the query "air-pollution". For this query, in the environmental semantic space, the document [01128042] is in the ranking 157th, that is, it is not selected in the high ranking. However, in the integrated environmental-medical semantic space, this document is in the ranking 4th. This result shows that this document can be selected in the integrated semantic space. The metadata for this document are "SPM, diesel, smoke, contamination, contaminant, warming, sea, environmental quality, environment agency, a standard, light oil, health, healthy damage, pollution, Oxide, The automobile NOx law, air pollution, Nitrogen, Nitrogen Oxide, Nitrogen dioxide, Exhaust gas, emission control, Exhaust, floating particle-like substance, Sulfur". The subset of the metadata "SPM, diesel, smoke-contamination, contaminant, warming, air pollution, Nitrogen oxide, Exhaust gas" of the metadata is highly related to "air pollution" in the environmental field, and the other subset of the metadata "health, healthy damage, pollution" are highly related to "air pollution" in the medical field. That is, this document is related to "air pollution" in the integrated environmental-medical semantic space, rather than in the single semantic spaces. This result shows that our method makes it possible to acquire information which is related to several fields and is not extracted in the semantic space created for a single field. Unlike our method, other methods using semantic spaces cannot extracted this kind of documents.

Similarly, Table 4 shows the retrieval results to the query "Environmental hormones". For this query, in the environmental semantic space, the document [000828030] is in the ranking 346th, that is, it is not selected in the high ranking. However, in the integrated environmental-medical semantic space, this document is in the ranking 9th. This result shows that our method realizes acquisition of information which is related to several fields and is not extracted in the semantic space created for a single field.

The metadata for this document are "SPM, cancer, Diesel, contamination, Greenhouse effect, sea, environmental standards, Environment Agency, a standard, Light oil, health, healthy damage, pollution, Oxide, automobile NOx law, the measure against an automobile exhaust gas, petroleum, air pollution, Carbon, Nitrogen, Nitrogen oxide, Copper, Carbon dioxide, Nitrogen dioxide, Exhaust gas, Emission control, the amount of discharge, Carcinogenic, Floating particle-like substance, fog, Sulfur".

In this document, the subset of the metadata "cancer, health, healthy damage, pollution, Carcinogenic" is highly related to "Environmental hormones" in the medical field. This document is not extracted in the environmental semantic space because the metadata are not highly related to "Environmental hormones" in the space, but it is extracted in the integrated environmental-medical semantic space.

Table 5 shows the retrieval results to the query "rehabilitation". For this query, in the medical semantic space, the document [980924091] is in the ranking 193th, that is, it is not selected in the high ranking. However, in the integrated environmental-medical semantic space, this document is in the ranking 5th. The metadata of this document

Table 3: The result for the query “Air pollution”

rank	Integrated Space	Environmental Space	Medical Space
1	001217119 - 0.757872	000625162 - 0.395851	971004314 - 0.532862
2	000826025 - 0.746731	000622083 - 0.365665	980514260 - 0.495517
3	990506132 - 0.746605	000120132 - 0.353257	980508179 - 0.483583
4	001128042 - 0.732719	000119220 - 0.349681	980606209 - 0.461891
5	980617296 - 0.729570	980903276 - 0.341292	971225110 - 0.455851
6	001106140 - 0.722226	980417025 - 0.324443	980417168 - 0.454193
7	980822158 - 0.721012	971015081 - 0.318102	980417134 - 0.428897
8	001015134 - 0.719486	001220024 - 0.313622	981223001 - 0.427919
9	000828030 - 0.718638	001201344 - 0.312631	981207252 - 0.407963
10	990630163 - 0.715068	000201030 - 0.312409	000503002 - 0.367019
-	-	-	-
155	990128019 - 0.550391	000630141 - 0.166195	980731226 - 0.202048
156	990705164 - 0.549977	990523208 - 0.164348	000603039 - 0.201249
157	000317080 - 0.549225	001128042 - 0.163946	980318249 - 0.201079
158	990817003 - 0.548865	000914166 - 0.161578	980609034 - 0.200232
159	990430215 - 0.548798	990409037 - 0.160712	980226112 - 0.199906
160	000916016 - 0.548206	000430155 - 0.160669	980108022 - 0.199830
-	-	-	-
495	981008357 - 0.240430	980421066 - 0.074441	981030300 - 0.112964
496	971023101 - 0.238382	001115316 - 0.073632	981105322 - 0.111208
497	000718090 - 0.237932	000522195 - 0.073462	990320218 - 0.111173
498	000820076 - 0.236857	980124137 - 0.072628	980518076 - 0.105826
499	980514260 - 0.236689	000815116 - 0.071398	980702379 - 0.103438
500	980606209 - 0.231906	001217119 - 0.070788	000519231 - 0.102851
-	-		
995	980122259 - 0.000000		
996	980119050 - 0.000000		
997	990129172 - 0.000000		
998	981201341 - 0.000000		
999	980320232 - 0.000000		
1000	990328037 - 0.000000		

Table 4: The result for the query “Environmental hormones”

rank	Integrated Space	Environmental Space	Medical Space
1	001217119 - 0.608303	990622146 - 0.490062	971004314 - 0.833046
2	990506132 - 0.602118	991209025 - 0.488231	990207146 - 0.805216
3	000826025 - 0.597233	001008152 - 0.482937	001203184 - 0.798642
4	001128042 - 0.588013	001219087 - 0.479456	980226374 - 0.796882
5	980617296 - 0.588009	990319268 - 0.471976	000228033 - 0.796095
6	001015134 - 0.583953	000917206 - 0.462762	980315082 - 0.789393
7	980822158 - 0.580817	000420149 - 0.459485	001018015 - 0.788370
8	001106140 - 0.579942	000724149 - 0.458495	980827211 - 0.784964
9	000828030 - 0.579025	000308112 - 0.455430	000208028 - 0.783548
10	990630163 - 0.575484	991008090 - 0.450773	000111020 - 0.783444
-	-	-	-
345	991002024 - 0.397573	990529311 - 0.165141	001023062 - 0.676683
346	990206130 - 0.396505	000828030 - 0.164548	981216292 - 0.676617
347	971023101 - 0.396302	001217119 - 0.164532	990403229 - 0.676071
348	980302004 - 0.396166	001112159 - 0.164246	971120001 - 0.676011
349	980904187 - 0.395591	001213222 - 0.164041	980518076 - 0.675793
350	990708027 - 0.395425	001015140 - 0.163888	001128142 - 0.675600
-	-	-	-
495	000827163 - 0.299448	990429123 - 0.108627	980404225 - 0.600715
496	980417363 - 0.299043	000615024 - 0.106554	980926161 - 0.598681
497	970522131 - 0.297935	990504190 - 0.105574	981211355 - 0.590334
498	990212137 - 0.297796	000622083 - 0.105187	980129109 - 0.585872
499	000730157 - 0.297184	980110032 - 0.099752	980926176 - 0.562776
500	000329078 - 0.296400	980307043 - 0.097569	000424258 - 0.545890
-	-		
995	000929221 - 0.053021		
996	990212158 - 0.052136		
997	990608208 - 0.051044		
998	001202168 - 0.050939		
999	000901028 - 0.047600		
1000	000518111 - 0.047534		

Table 5: The result for the query “Rehabilitation”

rank	Integrated Space	Environmental Space	Medical Space
1	000212252 - 0.343179		000212252 - 0.376446
2	990608334 - 0.306642		001217040 - 0.299593
3	980731226 - 0.305199		980731226 - 0.299593
4	001217040 - 0.305199		990510190 - 0.289963
5	980924091 - 0.304567		990129248 - 0.288849
6	990201257 - 0.300692		000503002 - 0.287556
7	001216154 - 0.295039		990101224 - 0.284870
8	980123363 - 0.292607		980713306 - 0.282685
9	990101224 - 0.292605		000721226 - 0.282491
10	001206315 - 0.289999		000727218 - 0.282465
-	-		-
190	990319234 - 0.223382		001102187 - 0.233054
191	990429202 - 0.223382		971228015 - 0.232904
192	980803175 - 0.223363		981016057 - 0.232896
193	980212131 - 0.222834		980924091 - 0.232876
194	000408035 - 0.222228		001023282 - 0.232871
195	000131170 - 0.222139		000914216 - 0.232793
-	-		-
495	990515229 - 0.162435		981007173 - 0.157683
496	001117222 - 0.161544		980421173 - 0.153499
497	001203184 - 0.161426		970919029 - 0.152549
498	000611151 - 0.161363		000402194 - 0.152527
499	980209031 - 0.160666		001114034 - 0.151237
500	980109222 - 0.160655		980801168 - 0.146089
-	-		
995	000818040 - 0.093834		
996	980901233 - 0.093075		
997	000713104 - 0.092315		
998	001023061 - 0.091412		
999	990429041 - 0.089518		
1000	980824168 - 0.087680		

are "MRI, epilepsy, paralysis, rehabilitation, disturbance of consciousness, hepatic cirrhosis, trachea, thrombus, hypercholesterolemia, hypertension, hyperlipemia, bleeding, tongue, apoplexy, position, bowel, hypotension, diabetes, head, arteriosclerosis, urine, Brain, cerebral thrombosis, stroke, Lungs, pneumonia, attack, hospital, pulse". These metadata are related to both environmental and medical fields. This result shows that our method makes it possible to extract information which is related to various fields.

These experimental results have shown that our semantic space integration method realizes a new semantic information acquisition and retrieval environment. This method can be applied repeatedly to integration of various semantic spaces created in different research fields.

5.3 Experiment-3

In Experiment-3, we evaluate the ability in terms of field independency in the integrated semantic space.

5.3.1 Evaluation method

As retrieval candidate documents for the integrated semantic space, we mapped 1000 documents (newspaper articles) into the space, where the 500 documents are from the environmental field and the other 500 documents are from the medical field.

Queries are classified into three kinds: (1) queries related to the environmental field, (2) queries related to the medical field, and (3)queries related to both environmental and medical fields.

For each query, query results are represented in the ranking for documents. Points are given to each document in the top 100, as 100 points for the 1st ranking document, 99 points for 2nd ranking, - - -, and 1 point for the 100th ranking.

For the queries related to the environmental field, it is expected that the total points for the documents related to the environmental field should be high in the query result. The experimental results are shown in Fig. 10, 11 and 12.

5.3.2 Experimental results

For the queries related to the environmental field, the documents highly related to the environmental field have been extracted in high ranking, and for the queries related to the medical field, the documents highly related to the medical field have been extracted in high ranking. Furthermore, for the queries related to both environmental and medical fields, the documents have been extracted from both environmental and medical fields.

Those experimental results show that the integrated semantic space has ability for realizing field independency, that is, for field-specific queries, documents related to the field itself can be extracted sharply. Furthermore, for queries related to various fields, our integrated semantic space realizes information acquisition from various fields.

Our experimental study has shown that our semantic space integration method and integrated semantic space realize new semantic space creation for integrating various research fields.

6 Conclusion

Our metadatabase system has been designed for dealing with semantic search space integration between heterogeneous databases in a multidatabase environment. In this system, the machinery for semantic associative search is realized as the essential semantic search system for extracting the semantically related information in a multidatabase environment.

In this paper, we have also presented the applicability of the semantic space integration method to the actual multidatabases for environmental and medical information. We have shown several experimental results which have been obtained by semantic search for two different databases in a multidatabase environment. In this experimental study, we have implemented an actual system for environmental and medical fields which are becoming important in the world-wide areas and clarified the feasibility and effectiveness of the metadatabase system in the actual multidatabases.

As the future work, we will extend the current integrated semantic space to include various databases in heterogeneous research fields. Furthermore, we will integrate the

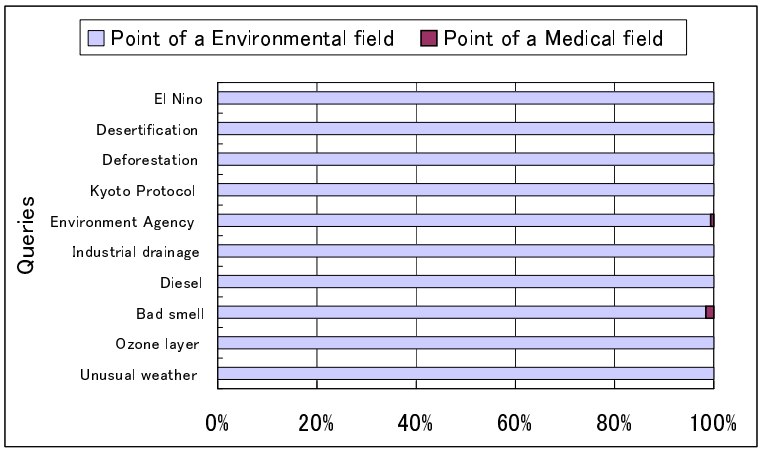

Fig.10: The result of queries related to the Environmental field

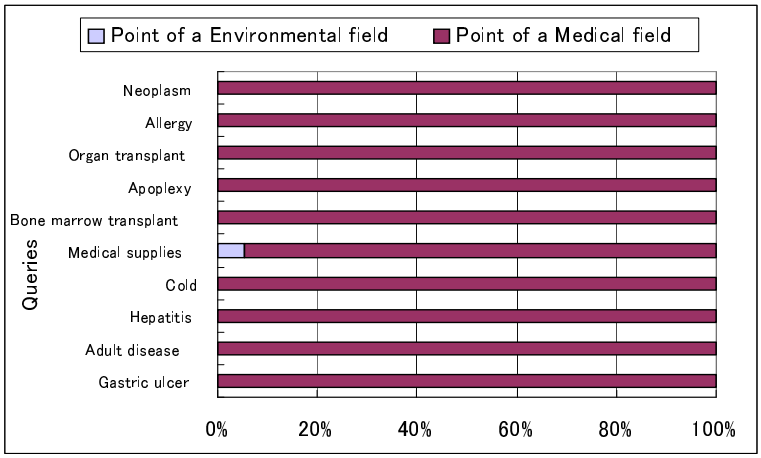

Fig.11: The result of queries related to the Medical field

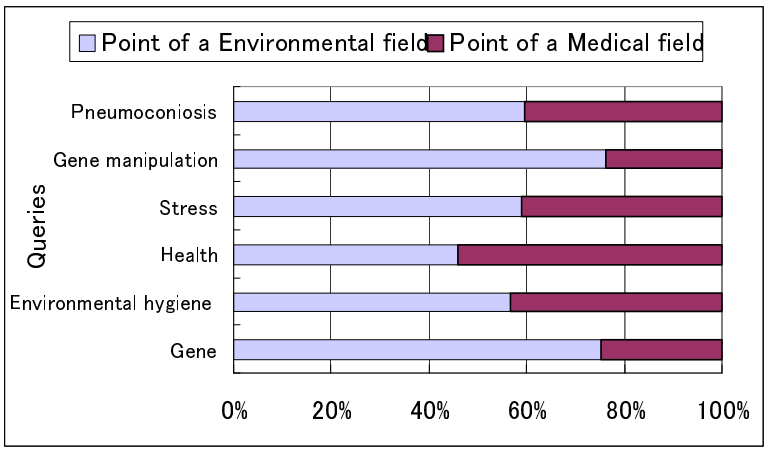

Fig.12: The result of queries related to both environmental and medical fields

semantic associative search system with the multimedia database systems[3, 4] in a distributed and network database system to realize multimedia databases for various research fields.

References

[1] Bright, M.W., Hurson, A.R., and Pakzad, S.H.(1992), "A Taxonomy and Current Issues in Multidatabase System," *IEEE Computer, Vol.25, No.3*, pp.50-59.

[2] Kitagawa, T. and Kiyoki, Y., *"The mathematical model of meaning and its application to multidatabase systems,"* Proceedings of 3rd IEEE International Workshop on Research Issues on Data Engineering: Interoperability in Multidatabase Systems, pp.130-135, April 1993.

[3] Kiyoki, Y. and Kitagawa, T., *"A metadatabase system for supporting semantic interoperability in multidatabases,"* Information Modelling and Knowledge Bases, IOS Press, Vol. V, pp. 287-298, 1994.

[4] Kiyoki, Y., Kitagawa, T., and Hayama,T., *"A metadatabase system for semantic image search by a mathematical model of meaning,"* ACM SIGMOD Record, (refereed as the invited paper for special issue on metadata for digital media), Vol.23, No. 4, pp.34-41, Dec. 1994.

[5] Kiyoki, Y., Kitagawa, T., and Hayama,T.,"*A metadatabase system for semantic image search by a mathematical model of meaning,"* Multimedia Data Management – using metadata to integrate and apply digital media –, McGrawHill(book), A. Sheth and W. Klas(editors), Chapter 7, 1998.

[6] Kiyoki, Y., Kitagawa, T. and Hitomi, Y., *"A fundamental framework for realizing semantic interoperability in a multidatabase environment,"* International Journal of Integrated Computer-Aided Engineering, Vol.2, No.1(Special Issue on Multidatabase and Interoperable Systems), pp.3-20, John Wiley & Sons, Jan. 1995.

[7] Kiyoki, Y. and Kitagawa, T., *"A semantic associative search method for knowledge acquisition,"* Information Modelling and Knowledge Bases (IOS Press), Vol. VI, pp.121-130, 1995.

[8] Kiyoki, Y. and Kitagawa, T. , *"Application of a Semantic Associative Search Method to Multidatabases for Environmental Information,"* Information Modelling and Knowledge Bases (IOS Press), Vol. XI, May, 1999.

[9] Larson, J.A., Navathe, S.B., Elmasri, R.(1989), *"A theory of attribute equivalence in database with application to schema integration,"* IEEE Transaction on Software Engineering, Vol.15, No.4, pp.449-463, 1989.

[10] Litwin, W., Mark, L., and Roussopoulos, N.(1990), *"Interoperability of Multiple Autonomous Databases,"* ACM Comp. Surveys, Vol.22, No.3, pp.267-293., 1990.

[11] *"Longman Dictionary of Contemporary English,"* Longman, 1987.

[12] Ogden, C.K.(1940), *"The General Basic English Dictionary,"* Evans Brothers Limited, 1940.

[13] Sheth, A. and Larson, J.A.(1990), *"Federated database systems for managing distributed, heterogeneous, and autonomous databases,"* ACM Computing Surveys, Vol.22, No.3, pp.183-236, 1990.

[14] Sheth, A. and Kashyap, V.(1992), *"So far (schematically) yet so near (semantically),"* Proc. IFIP TC2/WG2.6 Conf. on Semantics of Interoperable Database Systems, pp.1-30.

[15] Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A., *"Indexing by latent semantic analysis,"* Journal of the Society for Information Science, vol.41, no.6, 391-407, 1990.

[16] Dumais, S. T., Furnas, G. W., Landauer, T. K., and Deerwester, S., *"Using latent semantic analysis to improve information retrieval,"* Proc. CHI'88: Conference on Human Factors in Computing, New York: ACM, 281-285. p

[17] *"Dictionary for Environmental Problem Information, "* Nichigai Associates, Tokyo, 477p., 2001.

[18] Ueda, T., *"Handy Dictionary for Environmental vocabularies, "* Kyoritsu, Tokyo, 332p., 2000.

[19] Araki, S., *"Environmental Scientific Dictionary, "* Tokyo-Kagaku-Dojin, 1015p., 1985.

[20] Goto, C., *"Medical Dictionary, "* Version 2, Ishi-yaku, Tokyo, 1999.

[21] Hino, S., *"Medical Dictionary, "* Igaku-Shoin, Tokyo, 1107p., 1992.