Generation of Location-Related Knowledge from Web Contents

Yahiko Kambayashi, Ryong Lee and Taro Tezuka
Department of Social Informatics, Kyoto University
Sakyo, Kyoto 606, JAPAN
{yahiko, ryong, tezuka}@db.soc.i.kyoto-u.ac.jp

Abstract

The Web can be considered to be a large storage of various kinds of knowledge. How to use web contents is a very important and interesting problem. In this paper we will discuss on the generation of knowledge related to GISs(Geographic Information Systems). Since the real world and the world people recognize are different, we need to construct the latter world, in order to process queries as intended by users.

For such a purpose we need to develop efficient algorithms to obtain useful knowledge. Word frequency count is a very simple and computationally efficient, but still powerful. We have developed algorithms for (1) Landmark Identification, (2) Landmark Influence, (3) Landmark Interaction, and (4) Landmark Characterization, based on the word count method. Here, landmark is a name of places/buildings, for which people think it is important. It is a base of constructing the world people recognizes. After finding landmarks we need to consider problems (2)-(4). Some landmark has very strong influence and computation of strength of influence is handled by (2). If there are several landmarks located closely, we need to consider the interaction problem as shown by (3). A landmark can be a station, a temple, a public building or a crossing of streets, etc.

By analysing the usage of words for landmarks, we can show the characteristics of each kind of landmarks. If we use association rules, we can get better knowledge although to derive rules is time consuming $O(n^2)$ when n is the number of words. We can restrict the candidate words by analysing web contents. By restricting words like landmarks and typical characteristic words, useful association rules can be derived. We have developed a system called KyotoSEARCH, which has a unique user interface utilizing association rules for navigation purposes. The rules are combined with map interface and a related URL list. To get knowledge from web contents we have collected 2 million web pages, each of which is related to Kyoto. We have not developed algorithms to derive application dependent knowledge, but we believe such extension is not difficult. Although experiments are performed by web pages written in Japanese, the results in this paper are language independent.

1. Introduction

By the recent explosive growth of the web, it becomes important to retrieve necessary knowledge from the web. Web is a great source of dynamically changing knowledge. It is difficult to get appropriate results by applying just one query, we need to improve queries by analysing the result of the previous query[13]. By the analysis of the result, the users will know the characteristics of the data stored in the web.

In the case of GIS applications, query results are influenced by how people see the world. That is, the conceptual world is not equivalent to the real world. Instead of analysing each result, it will be very much useful to form appropriate queries if we know the characteristics of the conceptual world, since the conceptual world is shared by applications.

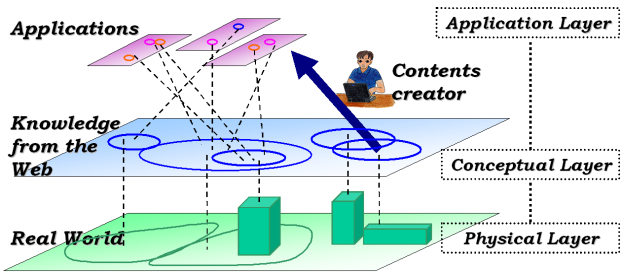

Fig.1 Three-Layer Model

Fig.1 shows the three-layer model to be used in this paper, which is similar to the three-layer model for database systems. The physical layer corresponds to the real world data. The conceptual layer corresponds

to the way people recognize the world. For example, some building is more important than others. There are several versions for the applications depending on characteristics of the applications. The conceptual world will be modified by the application. For example, sight-seeing spots are important for travellers, but shops and schools are more important for local residents.

The query modification approach discussed previously shows only a port of the application world as a result of the query. Sometimes it will be required to know the application world to get appropriate results. It is too much time consuming if we accumulate the results of repeated queries.

The approach to be used in this paper is to construct the conceptual world and each application world using the data distributed in the web. For example, the frequency of appearance of a place name can be regarded as the importance of the place recognized by users. If two place names appear simultaneously in many cases, it can be concluded that people think these two places are strongly related. We will develop the methods to construct the conceptual world from the contents in the web.

For such purpose we have collected about 2 million web pages, each of which is related to Kyoto City in Japan. As we are living in Kyoto, it is rather easy to judge whether the conceptual world generated is correct or not. The following unary relations and binary relations are used.

(1) **Unary Relations**: The world frequency is used to find out the importance.

(2) **Binary Relations**: Co-occurrence of words and conventional association rules[1] are used to get relationships among terminologies.

In order to organize the conceptual world, we classified the world into geo-words (words related to locations) and non-geo-words(others). These will be shown by G-words and N-Words for short, respectively. Frequency of G-words shows the importance of the location, we will use this technique to find landmarks, which are the place names usually used to identify locations. Fig.2 shows relationships among words. Relationships are classified into the relationships between G-words, the relationships between G-words and N-words, and the relationships between N-words. We can further restrict the set to be used for N-words. For example, in order to identify the characteristics of landmarks, we can restrict the words to be used to describe landmark characteristics (see the dotted lines in Fig.2).

In this paper, we will first discuss problems related to landmarks. How to identify landmarks is called a "landmark identification problem". To compute the strengths of landmarks is called a "landmark inference problem". Mutual interaction of landmarks is also important. These problems are discussed in Section 2.

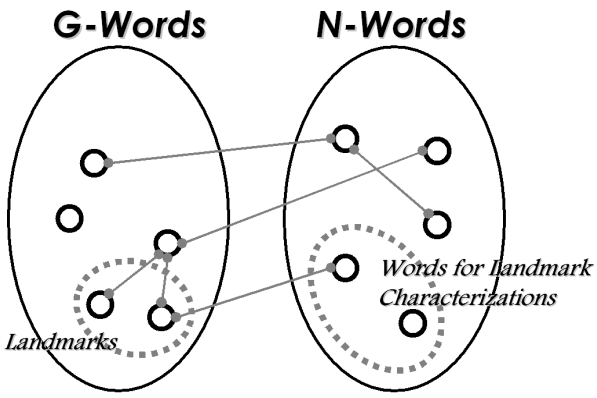

Fig.2 Relationships among words

One special case of utilization of relationships between G-words and N-words is to characterize landmarks. This problem is called a "landmark characterization problem", which is discussed in Section 3. All the algorithms above use the word count method, which is very much efficient. In general we need association rules, which require a time consuming process. We will restrict the computation time by selecting words to be used in the rules. An algorithms and some example of the results are shown in Section 4. Use of landmarks, association rules for integrated user interface for an advanced GIS is discussed in Section 5.

2. Landmark Identification and Landmark Influence

In computer systems the location is identified by latitude and longitude. People specify a location by street address or typical buildings. In U.S. address is very much systematic and it is rather easy to identify a location by street name and address number. In Kyoto the address number is not consecutively located. New number is assigned to a new house in the same area. The number shows building order of the first house at the location. In such a case specification of a location by typical spot, called landmark, is easy. How to find a landmark is shown below.

[Landmark Identification Problem]

(1) For each web page a list of G-words is generated. To identify G-words, we use a dictionary of G-words for Kyoto.

(2) Among G-words in the list, G-words corresponding to the address of the web site should be eliminated.

(3) There are several ways to determine weight of each G-word. One method is to put weight 1 to each G-word regardless the number of appearances. Another method is to put more weight to the G-word appearing frequently in the web page.

(4) Calculate the summation of the weights of each G-word for all the web pages to be considered. If weight 1 is assigned for each G-word in one web page, the weight obtained by this step is the number of web pages with this G-word.

(5) If the weight of a G-word computed by step (4) exceeds the pre-determined threshold value *d*, the G-word is selected as a landmark.

Apparently location names appearing in many web pages can be regarded as landmarks, since people think the names important and these names can be used for specifying locations. In order to obtain landmarks correctly, we need to analyze sentences in the web, so that the G-word is used to identify a location. We believe that simplified algorithm shown above is usually sufficient. The area influenced by each landmark is not the same. The number of appearance can be used to compute the influence.

Consider landmark A and count the number of phrase "near A". The owner of shop claiming "near A" thinks that A is the most influenced landmark among the landmarks near the shop.

As a summary, the method to calculate landmark influence is as follows:

[Landmark Influence Computation]

The influence value of each landmark is determined by the following (1), (2) or (3).

(1) Use the number of appearance of the landmark names

(2) Use the number of appearance of "near A", where A is the landmark to be considered.

(3) Instead of the number of obtained by (1) or (2) we can use root of one of the numbers.

The reason for (3) is as follows. If we measure the influence by the distance, then number of shops in distance *r* is one fourth of the number of shops in distance *2r*. That is, the root of the number of appearance of "near A" can be used to specify the influence.

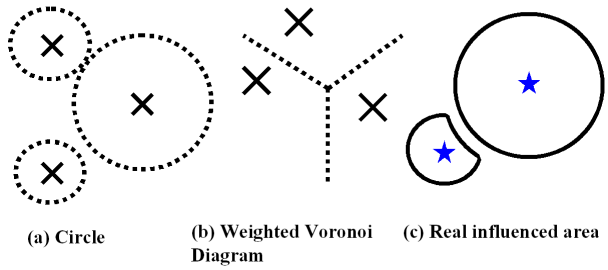

Fig.3 The areas influenced by landmarks

The area influenced by a landmark is distorted by the influence of other landmarks. In conventional GIS systems the area influenced by a landmark is specified by (a) a circle whose diameter is determined by the landmark influence number, (b) a Voronoi diagram with weighted importance. These cases are shown in Fig.3 (a) and (b). The real influence is shown in (c), which can be obtained by analyzing web data. In order to handle this kind of problems, we need to compute mutual influence of landmarks.

We will now discuss the interaction problem among landmarks. We assume the only landmarks located closely interact each other. The maximum distance is supported to be *d* (in the following example *d* is supposed to be 1,000 meters).

If there is an influential landmark B near the landmark A, it will give a great effect to A. We will define the effect of landmark B to A as follows:

(The weight of landmark B)/(the distance between landmark A and landmark B)^2

Fig.4 shows the computation of the effect for the landmark "Kyoto Station". Within distance *d* (1,000 meters) there are only two influenced landmarks, Kyoto Tower and Higashihonganji Temple. For simplicity, we will use the number of appearance in web pages as the weight of each landmark.

Kyoto Tower

Weight = 5500

Distance=298 meters

The effect of Kyoto Tower = 5500/298^2

= 0.0768

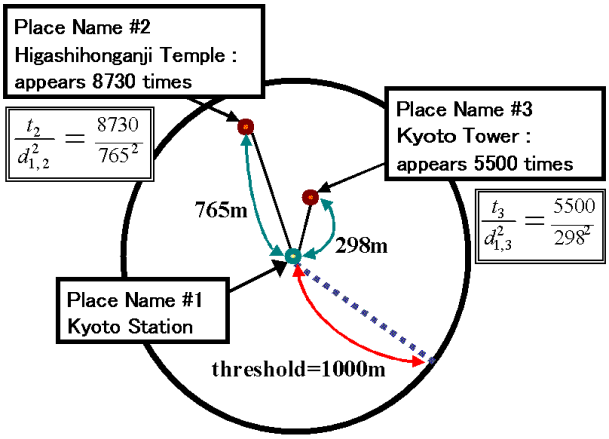

Fig.4 Landmark Interaction Problem

Higashihonganji Temple
Weight = 8730
Distance=765 meters
The effect of Kyoto Tower = $8730/765^2$
= 0.0149

Since there are only two landmarks within *d* (1000 meters), the amount of effect from other, landmarks to Kyoto Station is

0.0768+0.0149 = 0.0917

We can assume that if the effect of other landmarks to landmark A is low, the claim "near A" will be increased. Fig.5 shows the distribution of computation of the following pairs for typical landmarks in Kyoto.

(Effects from other landmarks, The number of "near A")

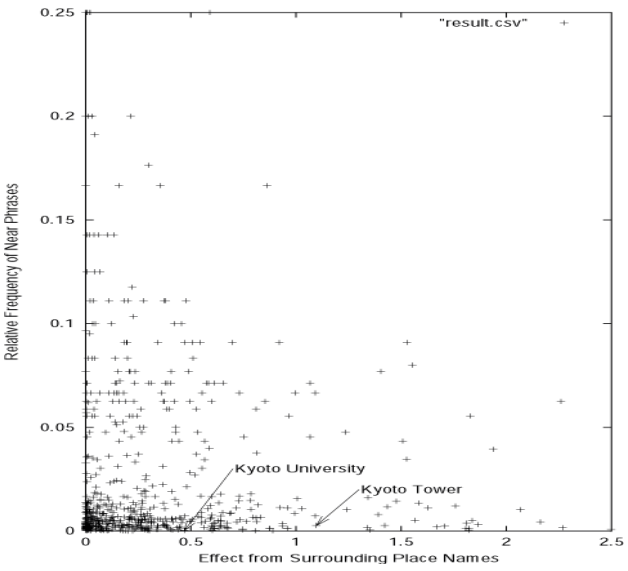

Fig.5 Distribution of the pair (Effect from other landmarks, the number of "near A")

The distribution shows the following tendency. If the number of "near A" is larger, the effects from other landmarks are smaller.

Note that the number of appearance of "Kyoto Tower" is very low, although the weight of "Kyoto Tower" is pretty high. As "Kyoto Station" is closely located, most neighbourhood of "Kyoto Tower" claims to be "near Kyoto Station". The reason why the number of "near Kyoto University" is small is different. As the university occupies a large area, it is hard to identify a unique location by "near Kyoto University".

To compensate the effect of the weight of landmark A, we can use the follow vales for y-axis.

(The number of "near A")/(weight of A)

Since the tendency is similar we will omit a graph for this case.

3. Characterization of Landmarks

In the previous section we have developed algorithms for problems related to landmarks. As all of them are related to the word count, only unary relations are used. In this section, a simple case of using binary relations is discussed. More general relations are discussed in Section 4.

Let *g* be a G-word and *m* be an N-word, general relationships between *g* and *m* are obtained by association rules. Here, we use word counts for g and *mg* (*mg* is a concatenation of m and g). For example, if *g* is "Kyoto Tower" and m is "near", *mg* is "near Kyoto Tower".

We will only consider landmark names as G-words and only a restricted set of N-Words as shown in Fig.2. The relationship will show the characteristics of the landmarks. The following words/phrases are used for N-Words.

near, around, in front of

For each landmark A, we counted the appearance of "near A", "around A", "in front of A", then the percentages are calculated. Landmarks are also classified into the following categories:

Temple, Crossing streets, Stations, Public buildings

Fig.6 shows the results of out experiments.

Around: For temples, crossings and public buildings, about 40% are "around A". For stations it is 60%.

In front of: For public buildings “in front of “ is used nearly 20%, and for temples 15%. There is very few usage of “in front of a station”, “in front of a crossing of streets”. These sentences are not well specified, since in front of or back of a station (crossing) cannot be precisely defined.

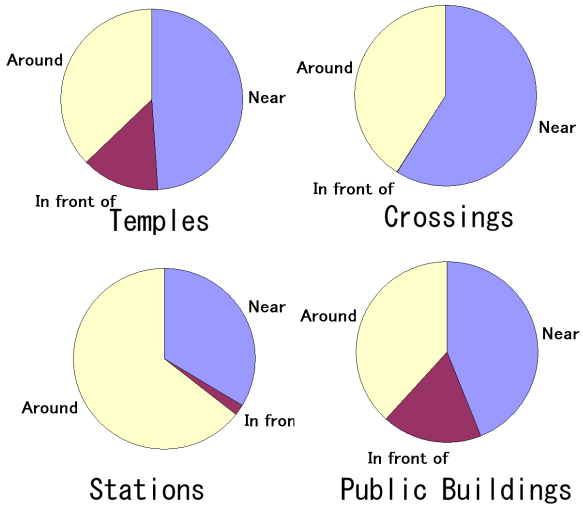

Fig.6 Characterization of Landmarks

Near: Near crossing of streets is rather commonly used. “Near Station” is not popular.

For landmark A if we count the number of appearance of “near A”, “around A”, “in front of A”, we may be able to know the category of the landmark.

In this section we used simple examples. In general, we may be able to categorize landmarks by N-Words appearing frequently with the landmark name.

4. Relationships among Words

We will use conventional association rules to derive relationships among locations and words. In web pages, there are many words and the number of relationship is n^2 where n is the number of words. Since it is impossible to compute relationships when n is large, reduction of the number of relationship is important.

(1) For G-words, we will only use landmarks.
(2) For N-Words, we have selected 185 words used for tours, festival, sports, museum, food, universities, traditions, etc.

The number of landmarks will change if we change the threshold value for the weights of landmarks. We can also select N-Words from the web. For restricted applications such as tour plans, we can select specific words for N-Words. Before describing the experiments, we will give formal descriptions.

[Web-based Association Rules]

Let $A = \{t, t_2, ..., t_m\}$ be a set of words. Let W be a set of web pages, where each w is a web page defined by a set of words in A. In order to reveal statistical significance, support is defined; a web page w is said to support a word t_i, if t_i is present in w. A web page w is said to support a subset $X \subset A$, if w support each $t_i \in X$. A set of words $X \subset A$ has a support s in W if $s\%$ of pages in W support X. On the basis of the support, association rule is defined: for a given set W of web pages, an association rule is an expression of the form $X \Rightarrow Y$, where X and Y are subsets of A and $X \Rightarrow Y$ holds with *confidence* c, if $c\%$ of pages in D that support X also support Y.

We can use the mining results to characterize each area/location. Some of the results are as follows:

[Characterization of a location]

(Shijo-Omiya is a downtown area in Kyoto City.)
Shijo-Omiya ⇒ gourmet | s=0.53%, c=78.32%
Shijo-Omiya ⇒ bar | s=0.53%, c=44.61%
Shijo-Omiya ⇒ noodle | s=0.53%, c=41.47%

There are gourmet, shops, bar and noodle restaurant at Shijo-Omiya, since it is in a downtown Kyoto.

(Ginkakuji is a famous temple. English translation is the Silver Pavilion.)
Ginkakuji ⇒ Tour | s=0.34%, c=30.62%
Ginkakuji ⇒ Architecture | s=0.34%, c=21.70%
Ginkakuji ⇒ Culture | s=0.34%, c=18.99%
Ginkakuji ⇒ History | s=0.23%, c=13.56%

Ginkakuji can be concluded as sight-seeing sports, architecture of building also attract attentions.

Kyoto University ⇒ University | s=1.38%, c=50.28%
Kyoto University ⇒ Education | s=1.38%, c=38.93%
Kyoto University ⇒ Culture | s=1.38%, c=26.03%

The result shows the characteristics of Kyoto University. Some of the results in opposite direction are as follows (Toji is also a well-known temple).

[Obtaining location names related to a given N-word]

noodle ⇒ Shijo-Omiya | s=0.53%, c=21.08%
temple ⇒ Toji | s=0.44%, c=14..86%
temple ⇒ Ginkakuji-temple | s=0.44%, c=5.55%

5. Use of Word Relationships for Advanced GIS User Interface

In the previous section we showed relationships of the following kinds:

G-word ⇒ *N-word*
N-word ⇒ *G-word*

Additionally, we can have the following relationships.

G-word ⇒ G-word
(related location is shown)

N-word ⇒ G-word
(related N-word is shown)

We have developed an advanced GIS system utilizing knowledge (relationships among words) extracted from the web.

For example, a foreigner plans to make sight-seeing of Kyoto City. S/he first would like to know the characteristics of Kyoto. The rule

'Kyoto'⇒'temple'

shows one of the characteristics of Kyoto is temples. By using the rules with temple in left side, we can identify famous temples. Successive application of the rules forms navigation of the city information. Based on the idea, we have developed an integrated geo-spatial web search system named KyotoSEARCH[7]. The system has an interface as shown in Fig.7, consisting of three components; Map, Keyword, URL-List Interfaces. The major functions of each interface are as follows:

a. **Map Interface** is prepared for specifying a location. The result of a query can be also shown on a map, which is easy to understand. Here we develop two kinds of maps: The left one is used for overview of the city with information of the number of web pages at each district, shown by colours. The right one is a detailed map at a specified location.

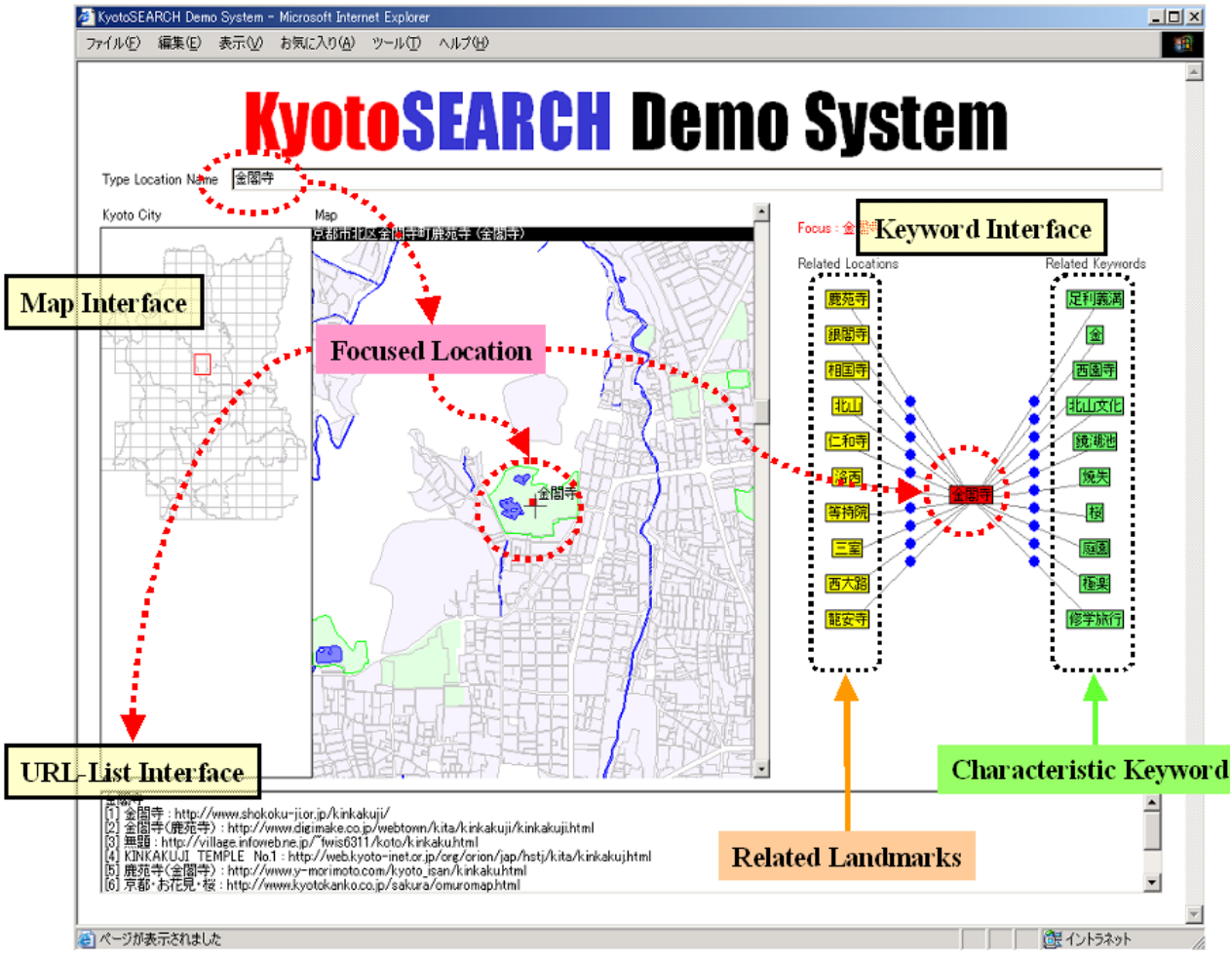

Fig.7 A Geo-spatial Web search System, KyotoSEARCH

b. **Keyword Interface** is used for navigating the association rules of G-words and N-words described in Section 4. In the centre of this interface, a focused keyword (a G-word or an N-word) is positioned. Its related G-words and N-words are placed around the focused keyword that is specified by users. If users click one of the related words, it becomes to a new focus, moves to control position, and shows rules related to new central word.

c. **URL-LIST Interface** is made for displaying a list of web pages relative to focused one in above Keyword Interface. The order of URLs is determined by the navigation history of the user.

In Fig.7, the major rules related to Kinkauji are shown. Related locations (G-words) are shown in the left side. Related N-words are shown in the right side. In the central word has changed, all the rules displayed will be changed accordingly. We can perform navigation using association rules in the system.

6. Conclusion

In this paper we have shown an advanced GIS system, using the knowledge obtained from the web contents. As the knowledge shows that how people see the world, we have to use the knowledge to properly answer a user's query. Some important knowledge can be obtained by simple word count operations. To get general knowledge, methods to obtain association rules can be applied. We believe that similar techniques can be applied various applications other than GIS. As we have applied our algorithms to web pages written in Japanese, there may be some language dependent nature. We believe, however, basic idea developed have can be applied to web pages in any language.

This work has been supported by 'Universal Design in Digital City' Project in CREST of JST (Japan Science and Technology Corporation).

References

[1] R. Agrawal and R. Srikant. "Fast algorithm for mining association rules," In Proc. of the 20th VLDB Conference, pages 487-499, Santiago, Chile, 1994.

[2] B. Bennet, ``Modal Logics for Qualitative Spatial Reasoning", Bulletin of the IGPL 3, 1996.

[3] M. J. Egenhofer and A. U. Frank, "Towards a Spatial Query Language: User Interface Considerations", Proc. of the 14th VLDB Conf., pp. 124-133, Los Angels, 1988.

[4] M. J. Egenhofer and D. M. Mark, ``Naive Geography", In A. U. Frank and W. Kuhn (Eds.), Spatial Information Theory: A Theoretical Basis for GIS, LNCS 988, pp.1-15, Springer, Berlin, 1995.

[5] J. Glasgow and A. Malton, ``A Semantics for Model-Based Spatial Reasoning", Mental Models in Discourse Processing and Reasoning, G. Rickheit and C. Habel (Eds.), pp. 259-297, Elsevier Science, 1999.

[6] R. Kasturi, R. Fernandez, M. L. Amlani and W. C. Feng, "Map Data Processing in Geographic Information Systems", Computer, Vol.22, No.12, pp. 10-21, 1989.

[7] R. Lee, H. Takakura, and Y. Kambayashi, "Visual Query Processing for GIS with Web Contents," Proc. of the 6th IFIP Working Conference on Visual Database Systems, May 29-31, 2002. (to appear)

[8] O. Lemon, ``Semantical Foundation of Spatial Logics", 5th Int. Conf. on Principles of Knowledge Representation and Reasoning, Morgan Kaufmann, 1996

[9] D. M. Mark and A. U. Frank, ``Experiential and Formal Models of Geographic Space", Environment and Planning B 23, pp. 3-24, 1995

[10] D. Papadias and T. Sellis, ``On the Qualitative Representation of Spatial Knowledge in 2D Space", VLDB Journal, Special Issue on Spatial Databases, Vol3(4), pp. 479-516, 1994.

[11] C. Parent, S. Spaccapietra and E. Zimanyi, "Spatio-temporal conceptual models: data structures + space + time", Proc. of the 7th Int. Symp. on Advances in Geographic Information Systems, pp. 26 - 33, Kansas City, 1999.

[12] D. A. Randell, Z. Cui and A. G. Cohn, ``A Spatial Logic based on Regions and Connection", 3rd Int. Conf. on Knowledge Representation and Reasoning, Morgan Kaufmann, 1992.

[13] G. Salton and C. Buckley. "Improving retrieval performance by relevance feedback," Journal of the American Society of Information Science, 41(4), pages 288-297, 1990.

[14] T. Tezuka, R. Lee, H. Takakura, and Y. Kambayashi, "Web-Based Inference Rules for Processing Conceptual Geographical Relationships," Proc. of the 2nd Int. Conf. on WISE Vol.II, The 1st Int. Workshop on Web Geographical Information Systems, Kyoto, pp.14-21, IEEE Computer Society Press, Dec. 2001.