VISUAL QUERY PROCESSING FOR GIS WITH WEB CONTENTS

Ryong Lee
Department of Social Informatics, Kyoto University
ryong@db.soc.i.kyoto-u.ac.jp

Hiroki Takakura
Data Processing Center, Kyoto University
takakura@rd.kudpc.kyoto-u.ac.jp

Yahiko Kambayashi
Department of Social Informatics, Kyoto University
yahiko@db.soc.i.kyoto-u.ac.jp

Abstract

In many geographic objects such as a travel planning, the use of web information is significantly increasing. For an efficient support of such work, it is very important to combine web information with map semantics. Current web systems usually do not support map semantics. Conversely, conventional Geographic Information Systems (GIS) do not utilize the web resources. The purpose of the research is as follows: (1) to get semantics from the web contents to realize advanced GIS functions on geographic web searches, and (2) to develop a user interface which can utilize web contents and map semantics in an effective integrating way. For such a purpose, we construct two map semantics about geographic characteristics and relationships available on the web. Utilizing semantics, we have developed a prototype system, KyotoSEARCH; its main function is to support users' information navigations among the web, the map and web-based geographic knowledge, in an integrated way.

Keywords: Map-based Web Search, Map Semantics, Geographic Web Search

X. Zhou et al. (eds.), *Visual and Multimedia Information Management*

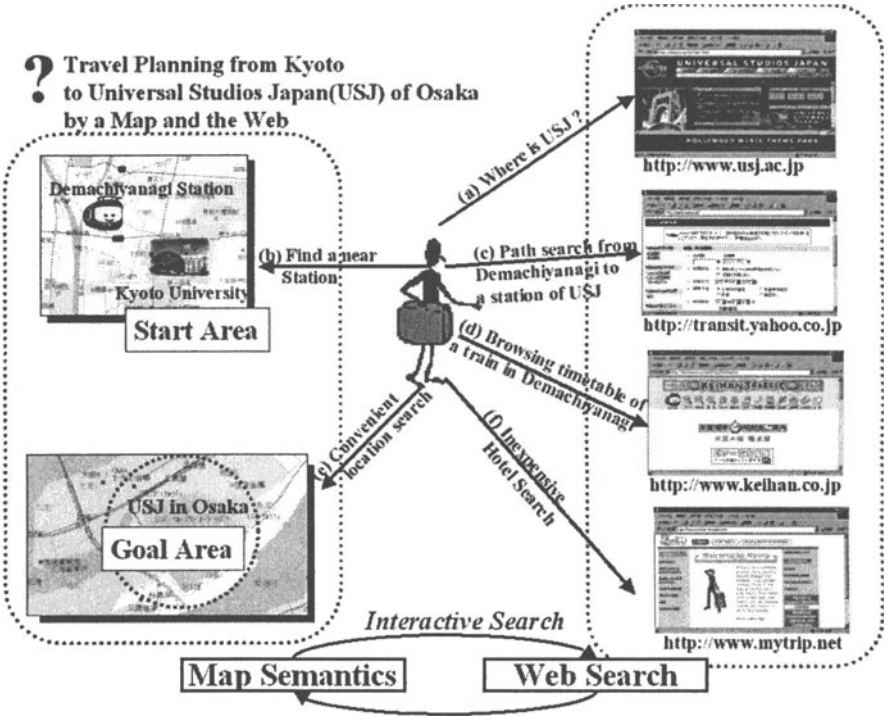

Figure 1. Geographical Information Search by a map and the web

1. INTRODUCTION

In the context of Geographical Information Systems(GIS), the current web resources should be another important database of human geographical information. In recent years, there are various kinds of significant efforts to integrate the web and geographical resources such as place names. Most of the efforts and possible extensions can be categorized as follows:

Indexing the Web by Relevant Locations

[Ding00; McCurley00; Arikawa00; Buyukkokten99]

By extracting place names from a page, a set of relevant geographic locations can be calculated. These locations represent the page's geographical coverage and relevancy. This will introduce new web classifying and indexing ways. We can use it for improving most the current web search engines that have been less focused on geography of the web.

Use of the Map as a User-Friendly Web Interface
[Lee00; Yates00; Hiramatsu01; McCurley00; Kumar99; BIGwhat; Mapion]
Instead of specifying locations by place names or latitude/longitude pairs, a user can select a location on the map precisely. In this case, other keywords should be specified separately, but it is possible to use geographical operations such as range and distance constraints.

Integrating of the Web Information with Map Semantics
If we can aggregate web resources highly related to a specific geographical location, it will be possible to perform spatial knowledge discovery on the web. That is, the web as a human geographical database will reveal unknown spatial knowledge. Then, it will be also used to improve web searches in geographic query processing.

The major objectives of this paper are as follows.

- *To realize an integrated system to advance GIS functions with the web*
- *To utilize the web as geographical knowledge base*

In order to describe problems of geographical web searches on the current web, let us consider a following scenario when we make a travel plan using the web.

A Motivated Scenario:
A foreign person who will participate in a Symposium at Kyoto University also would like go to Universal Studios Japan(USJ), that is a theme park in Osaka in Japan. We assume that she has only a map and a mobile computer connected to the web. First she will browse the USJ's web page as shown in Fig.1(a). Then, she can know the precise location of USJ. However, she wants more to know how to go there by train. To search for a route, the next query she posed is to browse a page about the train route search at a Yahoo! service page like Fig.1(b). This search, now, needs to be inputted 'starting station' and 'targeting station'. The latter one can be known by the USJ's page. For the determination of the 'starting station', she opens the map, and finds the nearest station 'Demachiyanagi' from 'Kyoto University'. Returned to the route search, she can now find a path which will be the best solution in conditions of charge, time, the number of transferring. In the next search, to find the timetable about a train of 'Demachiyanagi' station, she searches a page in the step of Fig.1(d). Furthermore, to reserve a hotel at a convenient place near the USJ, she will look for places around USJ on the map, and

found some hotels drawn by an image(Fig.1(e)). In order to compare price, facilities, etc. and to reserve one of them, she accesses to a hotel guide page.

As the above scenario shows, the web must be a useful resource for decisions of the planning which requires much of geographical knowledge with a map. Through her investigations with a map and the web, she could finish her preparations with much efforts and long time to this step. In the result, it is a very hard work, because the web and the map information are not integrated.

In order to utilize the web as a geographical knowledge base for advanced GIS, we focus on two kinds of important factors, geoword(place names, G) and non-geoword(N) founded in web pages(P). On the basis of the two kinds of word domains, we examine co-existence and association rules such as $G \rightarrow G$ *and* $G \rightarrow N$ by applying data mining methods. Here, for example, $G \rightarrow G$ shows an association rule for two geographical words(when W_1 exists there in many cases, W_2 exists in the same page). These relationships will derive a new semantic model for GIS such as **geographical characteristics and geographical relationships.** Moreover, we can benefit from utilizing these relationships in performing advanced geographic web search and web knowledge discovery.

The remainder of the paper is organized as follows. In Section 2, problems of conventional GIS's are discussed by the stand point described above. Section 3 describes how to compute associations and constraints of three domains(G, N, and P). Section 4 introduces a user-friendly comprehensive visual interface. It can extend GIS functionalities to Map-based Keyword Retrieval and Keyword-based Map Retrieval. In order to describe how to solve spatial queries efficiently on the domains, we discuss a web-based spatial query processing strategy in Section 5.

2. PROBLEMS OF CONVENTIONAL GEOGRAPHIC INFORMATION RETRIEVAL ON THE WEB

In solving spatial queries as the above example, searching for well-arranged tour guide web sites may be one solution today. However, as a generalized solution to these spatial queries with various purposes, web search engines should be integrated to GIS functions and resources.

All of these searches must be solved in two very fundamental domains; *spatial and non-spatial information domains.* Moreover, we need to refine each search results by applying *spatial constraints* such as region or distance, and *non-spatial constraints* such as term hierarchy. However,

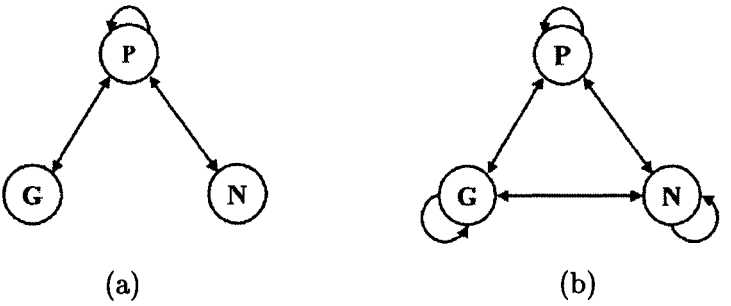

Figure 2. Relationships among P, G, and N, for conventional and advanced systems: (a) Conventional Web Information Systems, and (b) Relationships for Advanced GIS

most of current web search engines support only information navigations on domains of pages and related keywords. For complete navigations including geographical knowledge, the two domains, spatial and non-spatial information spaces should be strongly connected as shown in Figure 2. In the current Web, concepts in G(geoword) and N(non-geoword) are not directly connected with each other and itself as shown in Fig.2(a). We can say that they have some relationships (if they appear in the same web page). By analyzing web pages users can generate relationships between G and G' (other location names), between N and N' (other keywords), and between G and N. In the pairs, a geoword and a non-geoword can be related even if they do no appear in one particular page the user is interested in. However, it will give users an opportunity to know other interesting knowledge.

Generally, these kinds of relations can be a new semantics for GIS and Geographic Web Searches.

- *G-domain* has map semantics such as range or distance relationships in real worlds. In order to specify a geographical query and to display query result, map interface can be integrated to conventional web browsers.

- *N-domain* represents conceptual networks of terms which have been studied for a long time in textual processing study. It already has been constructed many terminology relationships such as similarity and term hierarchy. Languages have dynamic nature, there are also relationships among non-geoword(Ns) not contained in conventional dictionary such relationships can be found from the contents of web pages.

- *P-domain* has been constructed well-developed web search technologies in web search fields based on links and contents of the web.

By combining of these semantics, more powerful spatial knowledge supports are possible. This paper will construct the knowledge based on the association and constraints of the three domains. Comparable studies in information navigation is DualNAVI[Takano00]; it supports an information navigation on association of document and word space. Users can move from one document to another associative document by their link, and from one document to its most associative keyword. At the same way, movement from one keyword to another keyword or to the document space is possible. Our purpose is more general form to realize geographical information search by integrating web document space with map semantics.

3. CONSTRUCTION OF WEB-BASED MAP SEMANTICS

The term *Web Mining* has been used to refer to three kinds of data mining to Content, Usage, and Structure of the Web. The first one, on which we mainly focus in this paper, involves the discovery of meaningful knowledge from a large collection of primarily unstructured web data. This type of analysis is generally performed by means of interpreting statistical attributes of the discovered rules or patterns. In this paper, we exploit such discovery of the web in order to reveal the following geographical knowledge produced and shared by web users, where $G+$(or $N+$) shows a set consisting of elements in G(or N) respectively excluding empty sets.

Geographical Relationships : $G \rightarrow G+$

Geographical Characteristics : $G \rightarrow N+$

For example, results of most web search services about a location name 'Seoul in Korea'(G) in the end of May, 2002, will be shown many web pages extensively including related-location names ($G+$={'Niigata in Japan','Ulsan in Korea',...}) and characteristic words($N+$={'FIFA', 'World-Cup','Match Schedule','Team','Ticketing',...}), since the two cities take place '2002 FIFA World-Cup' together. Such relationships are very important at that moment and later its important will be decreased.

These kinds of knowledge extracted from the web are very different from those of conventional GIS based on the relational/object databases. Since the web space is constantly updating its contents in a large amount,

well-refined geographical knowledge of the Web can be a valuable source in geographical object applications. In the following subsection, we describe how to compute associations between geoword and non-geoword from web pages, and constraints in each domain for more efficient query processing.

Association Construction from Web Pages

The most straightforward and effect way in mining associations is to find the patterns which are relatively strong, i.e., which occur frequently together in most cases. In the data mining field, an association rule is a general form of dependency rule on transaction-based database; the rule has the form of "$W \rightarrow B$" (c%), explained as "if a pattern W appears in a transaction, there is c% possibility(confidence) that the pattern B holds in the same transaction", where W and B are a set of attribute values. In order to ensure that frequently encountered patterns is covered enough, the concept of the support of the rule was introduced, which is defined as the ratio that the pattern of W and B occurring together in the transactions vs. the total number of transactions in the database [Agrawal94].

Table 1. Fundamental Matrix

	nid_0	nid_1	nid_2	⋯	⋯	nid_m
pid_0	0	0	1	⋯	⋯	0
pid_1	0	0	1	⋯	⋯	0
pid_2	0	1	0	⋯	⋯	1
⋮	⋮	⋮	⋮	⋱	⋱	⋮
pid_n	1	1	0	⋯	⋯	0

(a) Matrix $M : P \times Noun$

	nid_0	nid_1	nid_2	⋯	⋯	nid_m
nid_0	20	2	3	⋯	⋯	9
nid_1	2	34	7	⋯	⋯	2
nid_2	3	7	16	⋯	⋯	3
⋮	⋮	⋮	⋮	⋱	⋱	⋮
nid_m	9	2	3	⋯	⋯	75

(b) $M^T M$

Table 2. Geo-Matrix

	gw_0	⋯	gw_k	ngw_0	⋯	ngw_l
pid_0	0	⋯	0	0	⋯	0
pid_1	1	⋯	0	1	⋯	0
pid_2	0	⋯	0	0	⋯	1
⋮	⋮	⋱	⋮	⋮	⋱	⋮
pid_n	0	⋯	0	1	⋯	0

(c) Matrix $G : P \times \{G, N\}$

	gw_0	⋯	gw_k	ngw_0	⋯	ngw_l
gw_0	10	⋯	3	32	⋯	5
⋮	⋮	⋱	⋮	⋮	⋱	⋮
gw_k	3	⋯	7	8	⋯	13
ngw_0	32	⋯	8	93	⋮	12
⋮	⋮	⋱	⋮	⋮	⋱	⋮
ngw_l	5	⋯	13	27	⋯	75

(d) $G^T G$

We mine the web by constructing a matrix M illustrated in Table 1., which defines the relationship between Page and Nouns. A row in (a) the matrix M represents noun-list appearing in a page pid_j. As the conventional mining work, the $page_{id}$ is corresponding with each shopping transaction, while the words of $page_{id}$ is a set of items included in each transaction. The co-citation matrix M^TM also can show the frequently associated noun-pairs. Here we consider a constraint that the occurrence of a noun in a page is counted just onetime for a brief description. Then, to find most relevant terms, the matrix M^TM has integer values, while M has binary values.

A mining rule that we are targeting is a rule of the form "$X \rightarrow Y$", where X and Y can be a set of $G+$ and $N+$; here, $G+$ is a set of geo-referential text(place names or geographical names), while $N+$ is a set of generic nouns excluding $G+$. For this, we introduce a matrix G in Table 2., which is made by distinguishing G from N. The co-citation matrix G^TG represents the three important relationships described in Figure 2, (i) $P \rightarrow \{P+, G+, N+\}$, (ii) $G \rightarrow \{P+, G+, N+\}$, (iii) $N \rightarrow \{P+, G+, N+\}$. Here, the relationship $P \rightarrow P+$ can be constructed from link structure among pages, i.e., $P+$ is a set of pages linked from page P.

In making above matrix, there are two way to process it from the web. One is for starting from aggregation of unknown data set of the web. In such case, we need to perform analysis work as following steps:

step 1. Extraction of *Page-Links, G, N* from contents of web pages $P \rightarrow \{P+, G+, N+\}$

step 2. Indexing for G, N search: Using G and N, we can construct index for web pages
1) $G+ \rightarrow P+$, 2) $N+ \rightarrow P+$

step 3. Association Construction: The following relationships are derived by the occurrence relationships of identical web pages.
1) $G+ \rightarrow G+$, 2) $G+ \rightarrow N+$, 3) $N+ \rightarrow G+$, 4) $N+ \rightarrow N+$

For information retrieval, words in $G+$ and $N+$ are determined, using index defined in step 2, and corresponding pages are obtained.

4. A WEB-BASED SPATIAL INFORMATION RETRIEVAL SYSTEM

In this section, we introduce a prototype system, KyotoSEARCH, to support information navigation among *G, N, and P.* The system has two main functions necessary to resolve the query about $G \rightarrow N$ and

$N \rightarrow G$. For that, new retrieval ways as Map-based Keyword Retrieval and Keyword-based Map Retrieval are introduced.

Our system has the following components as shown in Figure 3:

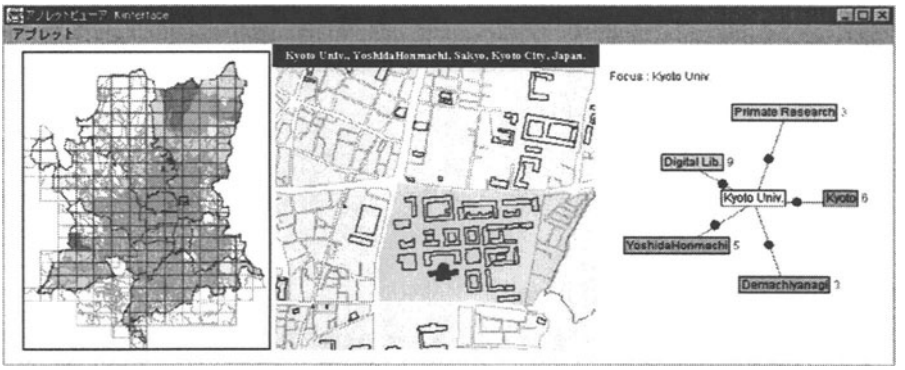

Figure 3. A user interface for KyotoSEARCH

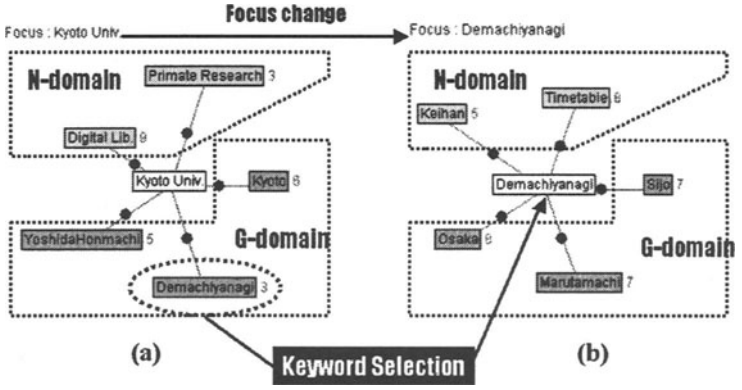

Figure 4. Knowledge Navigation on the Keyword Interface

- **Map Interface** is a great user interface to specify a location. The result of a query can be also shown on a map, which is easy to understand. Here we develop two maps: The left one is for show the number of web pages in each town. The right one is a detailed map at a specified location.
- **Keyword Interface** is used for exploiting the relationships of G and N introduced in the previous sections. In the center of this

interface, a focused keyword(G or N) is positioned. Its related $G+$ and $N+$ are placed around the focused keyword together with lines showing the semantic relationships(each relationship can be expressed by a label or a kind of line). If users click one of the related words, it becomes to a new focus, moves to center position, and re-shows its relatives as shown in Figure 4. In Fig.4(a), if a user selects "Demachiyanagi" as a new focused word, the graph will become the other shown in Fig.4(b).

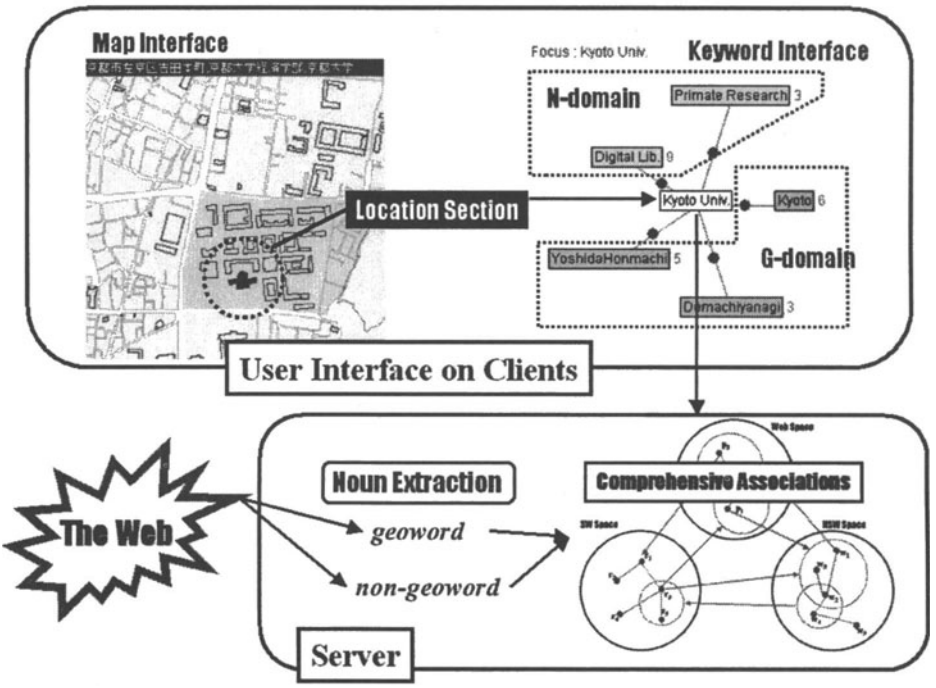

Figure 5. Map-based Keyword Retrieval

In addition, we made a URL-List Interface for displaying a list of web pages relative to focused one in above Keyword Interface. Users can browser most relative web sites by choosing one of them. In this paper, we do not describe specifically since our focus is on the above two interfaces. With above components, users can perform the following two retrieval operations alternatively.

Map-based Keyword Retrieval

By clicking a location to search on Map Interface, other interfaces are activated for receiving location key as shown in Figure 5. In Keyword Interface, the focused keyword changes into the received location keyword from the map, and shows a set of new relatives around it. URL-List Interface will show a list of web pages retrieved by the received location name. Users are then able to know relative information on them. This retrieval actually gives solutions to spatial queries starting from G, i.e. $G \rightarrow \{G+, N+, P+\}$.

Keyword-based Map Retrieval

A selection of one keyword from the relatives changes contents of Keyword Interface as shown in the above description. It also interacts with other interfaces; map interface shows the detailed map if the selected keyword is included in G. On the other hand, URL-List Interface shows a revised list related to the newly selected keyword.

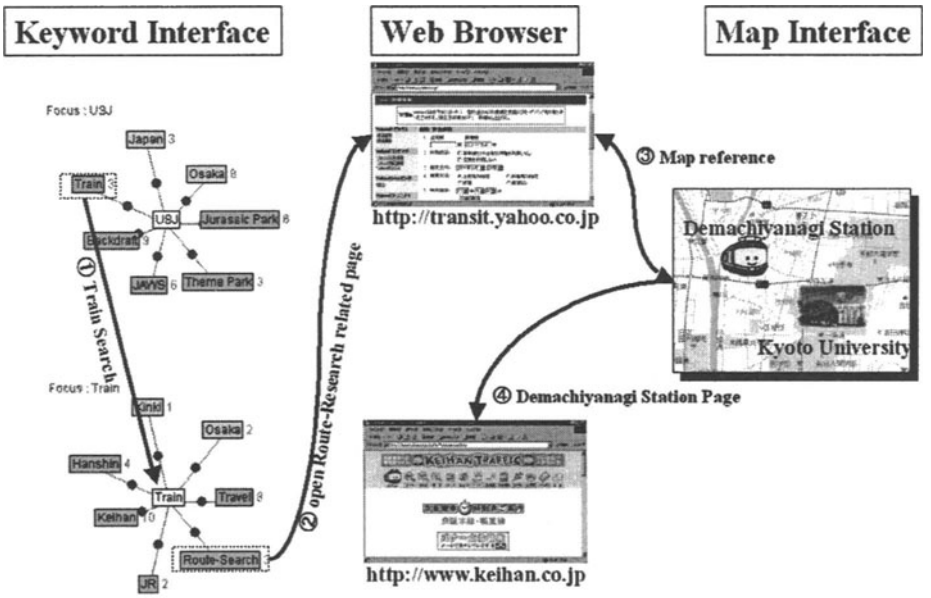

Figure 6. Protype system-based geographical knowledge navigation

We can now easily process the first scenario described in the introduction for travel planning as shown in Figure 6. The new travel plan process is as follows:

step 1. By specifying 'USJ' as the focused keyword, the corresponding map will be shown. (G='USJ' $\rightarrow$ K+={'Jurassic Park', 'JAWS', 'Train', ... }, G+={'Japan','Osaka'})

step 2. As one of the related words to the 'USJ' shown in Figure 6, we specify the 'train' as a new focused word. Then, as the result, new list of related words will be shown such as 'Route-Search', 'JR', 'Keihan', etc. (G='USJ' $\rightarrow$ K='Train' $\rightarrow$ P="http://transit.yahoo.co.jp')

step 3. By specifying Kyoto University as the focused keyword to view a map near the university, the most near station "Demachiyanagi" is found on the map. If it is selected, a web page related to "Demachiyanagi" will be retrieved. From this page, we get the information about the location of the station, timetable, etc.

step 4. By performing route search with "Demachiyanagi" as another web browser, we will get a list of most low-cost courses to 'USJ'.

5. DISCUSSIONS: EFFICIENT QUERY PROCESSING STRATEGIES

In the developing works and experiments using a large number of web resources, we faced to many difficult problems. First, the goal of our system was to make an integrated system to support multi-purposed geographic web search system. This made us suffer about how to reason the analysis results. Since the prototype system simply depends on high co-occurrence of terms without a consideration of any constraint or any pre-classification, it is very hard to evaluate the efficiency.

5.1. Constraints by Distance and Term Hierarchy

In order to improve the current search result, we discuss two constraints. One is spatial restriction to the results by applying actual distance constraint for filtering geowords. The other is for term hierarchy for filtering non-geoword.

For show what aspects will be improved with these constraints, we first consider the following example which has different characteristics from the previous examples.

Example) Simple Restaurant Search
One wish to find a good restaurant for dinner in a town the person does not know. This query now goes through the following steps:

step 1. Inquire into a location list that each location has many restaurants in the town

step 2. Choose a location from above location list

step 3. Request a restaurant type list in a selected location

step 4. Select one restaurant in above type list

step 5. Access to a web site of the selected one for more specific information

The result of above step 1 is desirable to show many related geographical regions on the map, so that one of the resulting regions need to be selected by users or automatically recommended ways. In step 3, she now need to know more specific information about what kinds of restaurant exist in the selected region. This question must be answered by using term hierarchy for 'restaurant'; it will have Chinese-, Italian-, France-, Japanese-, Korean restaurant, etc. as child terms. In step 4, one of the restaurants is selected by her favorites. She choose a 'Korean restaurant'. Finally, she access to a web page of the selected 'Korean restaurant' for looking at food menu and making reservation.

In resolving the above query, we first consider the following steps to solve the user query without constraints.

Formalized Simple "Restaurant" Search :

step 1. $LocationSearch('restaurant') : N \rightarrow G+$

step 2. $RestaurantList(G_{selected}) : G \rightarrow N+$

step 3. $PageSearch(restaurant_{selected}) : N \rightarrow P+$

In above search, the user needs to traverse on the whole information domains; First query is that what locations are related to "restaurant". It needs an association from N to $G+$ described in Section 2. Secondly the user asks what restaurants exist there. This involves $G \rightarrow N+$ search with selected G. By the result of this search, the user is given a restaurant list filtered by G and N conditions. With a final restaurant selection by the user, for the more specific information such opening/closing time, food menu, cost, etc., by use of $N \rightarrow P+$ association, the restaurant page is opened by web browser. Associations described in previous sections perform these transition from one domain to another. Through the traversal, users will get a comprehend spatial knowledge with non-spatial one based on the web resources. However, the simplified search will be resulted in large amount of low-quality answers to each query, since any constraints is yet considered. Therefore, we need to refine above query in the constraints about two domain G and N as follows:

Constraint-based "Restaurant" Search:

step 1. $LocationSearch('restaurant') : N \rightarrow G$

step 2. $Inside(1km, G_{list}) : G \rightarrow G$ **as Distance Restriction**

step 3. $RestaurantList(G_{selected}) : G \rightarrow N$

step 4. $ResturantType(N)$: $N \rightarrow N$ **as Categorization Restriction**

step 5. $PageSearch(restaurant_{selected}) : N \rightarrow P$

This search is now will show a list of the restaurants, restricted regions and categorized into Italian-, France-, Chinese restaurants, etc. These added constraints involves self-traverse in each domain of G and N which are possible at the strong connection we described in Section 2.

6. CONCLUSION

The association on the Keyword Interface is built on analysis of aggregated web pages from many web searches. The number of the aggregated web pages, we stored in local disks, is 2 millions as about 30GB data size. For extracting geowords and non-geowords from each page, we perform the morphological analysis, resulting in about 20GB output data size. As a future work, we will develop an efficient knowledge discovery method to handle such a large volume of web data.

The developed system, we described in Section 4, is implemented with Java Applet for using on the web after our further experimental work. This comprehensive interface realizes information navigation which we discussed at the strong connection in Section 2. We believe that many of difficult spatial queries in conventional GIS will be resolved by our proposed system by the following facts.

- Map interface is an intuitive interface for representing geographical regions of web resources and formulating user query in spatial domain.

- Keyword Interface utilizes a result from web contents mining in respect to G and N. It gives background and unknown knowledge to users by relations such geographical characteristics($G \rightarrow N$) and relationships($G \rightarrow G$).

Our another future work will focus on object-based knowledge discovery on the G-N-P graph, constructed by actual analysis work to millions of aggregated web pages.

ACKNOWLEDGEMENT

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

REFERENCES

R. Agrawal and R. Srikant. *Fast algorithms for mining associations*, VLDB1994, pp. 487-499, 1994.

M. Arikawa, K. Okamura. *Spatial Media Fusion Project*, Proc. of Kyoto International Conference on Digital Libraries: Research and Practice, pp.75-82, Nov. 2000.

O. Buyukkokten, J. Cho, H. Garcia-Molina, L. Gravano, and N. Shivakumar. *Exploiting geographical location information of web pages*, In Proc. of the ACM SIGMOD Workshop on the Web and Databases, WebDB, 1999.

J. Ding, L. Gravano,and N. Shivakumar. *Computing Geographical Scopes of Web Resources*, VLDB2000, pp.545-556, 2000.

K. Hiramatsu and T. Ishida. *An Augmented Web Space for Digital Cities*, IEEE/IPSJ Symposium on Applications and the Internet (SAINT-01), pp.105-112, 2001.

V. Kumar, A. Bugacov, M. Coutinho, and R. Neches. *Integrating Geographic Information Systems*, Spatial Digital Libraries and Information Spaces for conducting Humanitarian Assistance and Disaster Relief Operations in Urban Environments", ACMGIS, 1999.

F. Lee, S. Bressan, and B. C. Ooi. *Global Atlas: Calibrating and Indexing Document from Internet in the Cartographic Paradigm*, Internation Conference on Web Information Systems Engineering. Vol 1, pp. 117-124, 2000.

K.S. McCurley. *Geospatial Mapping and Navigation of the Web*, WWW10, 2000.

A. Takano, Y. Niwa, S. Nishioka, M. Iwayama, T. Hisamitsu, O. Imaichi, and H. Sakurai. *Associative Information Access Using DualNAVI*, Proc. of Kyoto International Conference on Digital Libraries: Research and Practice, pp.285-289, Nov. 2000.

J.D. Yates and X. Zhou. *Searching the Web Using a Map*, Proc. of the 1st International Conference on Web Information Systems Engineering, pp. 222-229, Jun., 2000.

BIGwhat. *http://www.bigwhat.com*

Mapion. *http://www.mapion.co.jp*