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

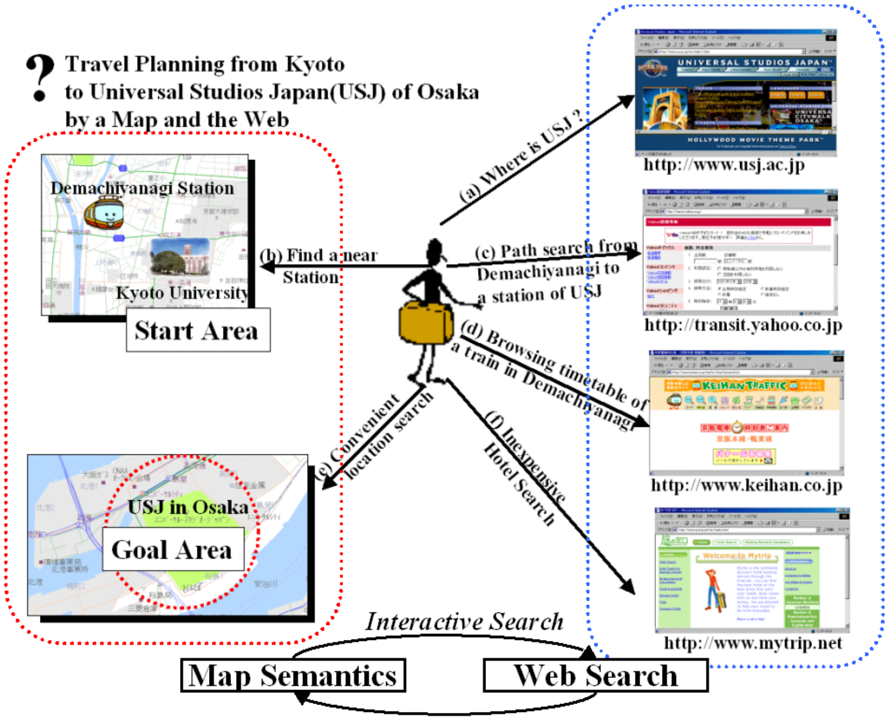
Travel Planning from Kyoto
to Universal Studios Japan(USJ) of Osaka
by a Map and the Web
Demachiyanagi Station
Kyoto University
Start Area
(b) Find a near Station
(a) Where is USJ ?
(c) Path search from Demachiyanagi to a station of USJ
(d) Browsing timetable of a train in Demachiyanagi
(e) Convenient location search
(f) Inexpensive Hotel Search
USJ in Osaka
Goal Area
http://www.usj.ac.jp
http://transit.yahoo.co.jp
http://www.keihan.co.jp
http://www.mytrip.net
Interactive Search
Map Semantics
Web Search

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, most of current web search engines support only information naviga-

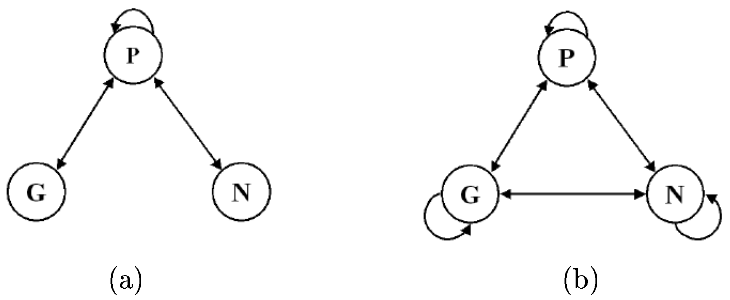

Figure 2. Relationships among P, G, and N, for conventional and advanced systems: (a) Conventional Web Information Systems, and (b) Relationships for Advanced GIS

tions 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	$\cdots$	$\cdots$	nid_m
pid_0	0	0	1	$\cdots$	$\cdots$	0
pid_1	0	0	1	$\cdots$	$\cdots$	0
pid_2	0	1	0	$\cdots$	$\cdots$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\ddots$	$\vdots$
pid_n	1	1	0	$\cdots$	$\cdots$	0

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

	nid_0	nid_1	nid_2	$\cdots$	$\cdots$	nid_m
nid_0	20	2	3	$\cdots$	$\cdots$	9
nid_1	2	34	7	$\cdots$	$\cdots$	2
nid_2	3	7	16	$\cdots$	$\cdots$	3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\ddots$	$\vdots$
nid_m	9	2	3	$\cdots$	$\cdots$	75

(b) $M^T M$

Table 2. Geo-Matrix

	gw_0	$\cdots$	gw_k	ngw_0	$\cdots$	ngw_l
pid_0	0	$\cdots$	0	0	$\cdots$	0
pid_1	1	$\cdots$	0	1	$\cdots$	0
pid_2	0	$\cdots$	0	0	$\cdots$	1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
pid_n	0	$\cdots$	0	1	$\cdots$	0

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

	gw_0	$\cdots$	gw_k	ngw_0	$\cdots$	ngw_l
gw_0	10	$\cdots$	3	32	$\cdots$	5
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
gw_k	3	$\cdots$	7	8	$\cdots$	13
ngw_0	32	$\cdots$	8	93	$\vdots$	12
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
ngw_l	5	$\cdots$	13	27	$\cdots$	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 georeferential 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.

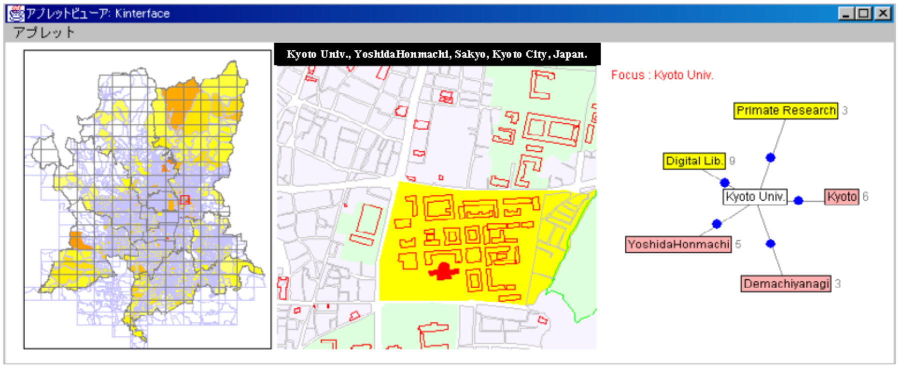

Figure 3. A user interface for KyotoSEARCH

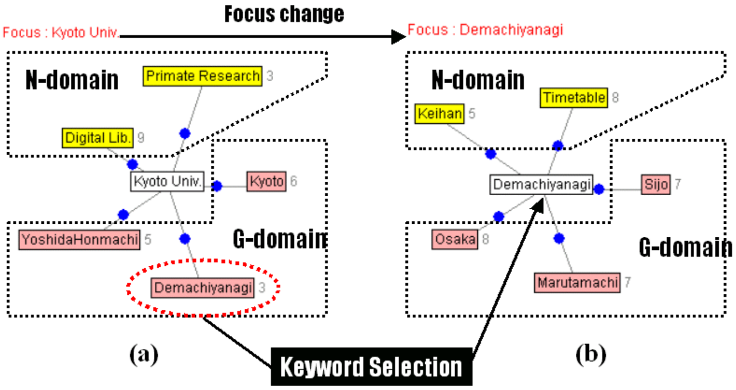

Figure 4. Knowledge Navigation on the Keyword Interface

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

- **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).

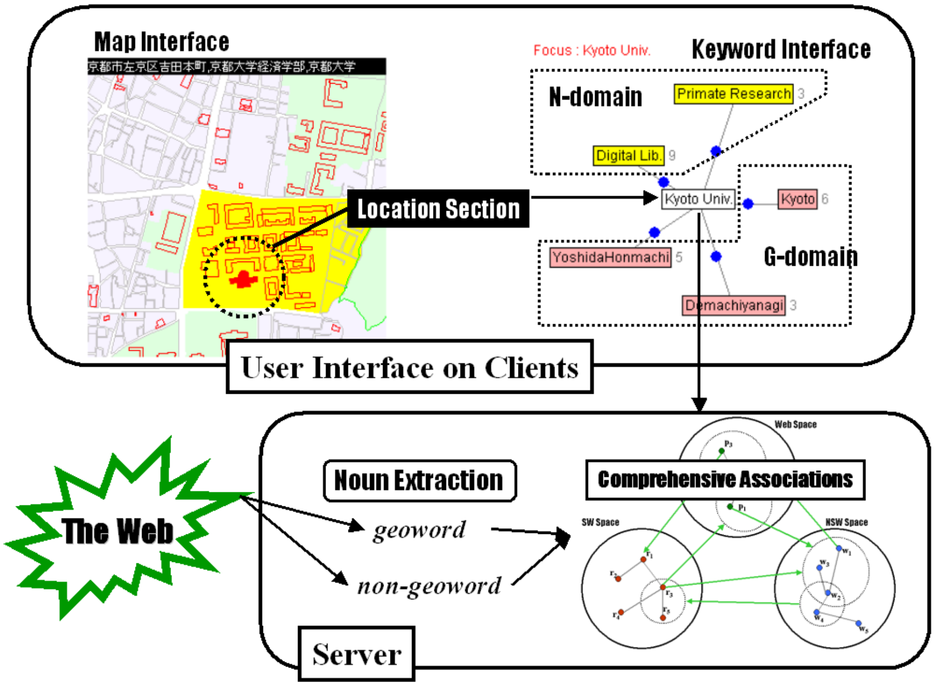

Figure 5. Map-based Keyword Retrieval

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