Toward Tighter Integration of Web Search with a Geographic Information System

Taro Tezuka
Kyoto University
Yoshida-honmachi, Sakyo-ku
Kyoto, Japan
tezuka@dl.kuis.kyoto-u.ac.jp

Takeshi Kurashima
Kyoto University
Yoshida-honmachi, Sakyo-ku
Kyoto, Japan
ktakeshi@dl.kuis.kyoto-u.ac.jp

Katsumi Tanaka
Kyoto University
Yoshida-honmachi, Sakyo-ku
Kyoto, Japan
tanaka@dl.kuis.kyoto-u.ac.jp

ABSTRACT

Integration of Web search with geographic information has recently attracted much attention. There are a number of local Web search systems enabling users to find location-specific Web content. In this paper, however, we point out that this integration is still at a superficial level. Most local Web search systems today only link local Web content to a map interface. They are extensions of a conventional stand-alone geographic information system (GIS), applied to a Web-based client-server architecture. In this paper, we discuss the directions available for tighter integration of Web search with a GIS, in terms of extraction, knowledge discovery, and presentation. We also describe implementations to support our argument that the integration must go beyond the simple map-and-hyperlink architecture.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Design

Keywords

Web-GIS integration, Local Web search, Web mining

1. INTRODUCTION

Local Web search is attracting much attention recently. Many practical applications have been implemented for searching Web content related to specific regions.

Major search engines and portal sites, such as Google, Yahoo!, Ask Jeeves, and MSN, provide local search services [32][33][34][35][36]. Most of these services, however, are based on what we call the *map-and-hyperlink architecture*. That is, the Web content and geographic information are weakly bonded: the Web pages only refer to geographic locations in their content.

A system based on the map-and-hyperlink architecture is illustrated in Figure 1. On one side of the interface, there is a map. On the other side, a list of snippets, the page title, and link URLs are presented. The user can click on the URL, the page title, or a marker on the map to see the linked Web content.

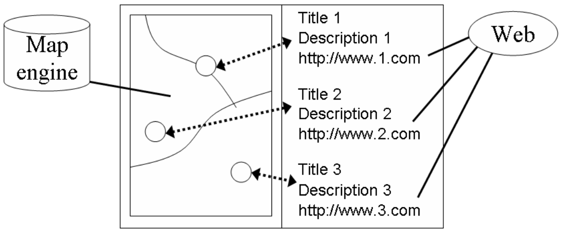


Figure 1: Map-and-hyperlink architecture

Although this architecture is very convenient for the user in many cases, it nevertheless has limited capability in other situations. For example, a user searching for information on restaurants while driving cannot pay much attention to a map interface, since it would take his or her attention away from driving. Another user might want a list of all the restaurants in a certain area. Showing a list of their URLs would not be very satisfactory if the list was extensive, since it would take much time to open all the pages and check their content. Summarizing the content would be more useful. A third user might want to learn about people's opinions of a region, according to what is being said on the Web. Such information is not provided by existing local Web search systems.

These services can be enabled through the tighter integration of Web search with a geographic information system (GIS). In this paper, we discuss the integration issues in terms of extraction, knowledge discovery, and presentation. We argue that the integration should go beyond the simple map-and-hyperlink architecture, and we support our argument by describing several implementations.

The rest of the paper is organized as follows. We discuss related work in the next section. In Section 3, we give an overview of the tighter integration of Web search with a GIS. In Section 4, we discuss the extraction of geographic information from the Web. Sections 5 and 6 discuss knowledge discovery and presentation, respectively. Finally, Section 7 is the conclusion.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

2. RELATED WORK

There are now a number of local Web search systems open to the public. Most of these are based on the map-and-hyperlink architecture, including those developed both as research projects and as commercial applications.

A typical example is the system developed by McCurley [1]. The system maps Web pages to geographic locations according to IP addresses, physical addresses, telephone numbers, and other information. The spatial browser consists of a map interface and a regular Web browser. When the user clicks on a point on the map interface, a menu appears, showing the URLs of pages related to that location on the map.

Larson discussed geographic information retrieval on the Internet [2]. He described the concept of a *hypermap*, which is a map version of hyperlinks; the maps and geographic content are provided on the Web and linked to each other in a similar manner as with hyperlinks. Plewe also described various means of providing geographic information over the Internet. Most of these means involve map interfaces [3].

Many search engines and portal sites now provide local Web search services. Google, Yahoo!, Ask Jeeves, and MSN, to name a few, have their own local search systems. These services are also based on the map-and-hyperlink architecture.

Google's local search service, Google Local, shows search results on a map interface [32]. Its geographic database contains the locations of shops, hotels, restaurants, and so forth. Unlike Google's original Web search engine, Google Local is heavily dependent on databases provided by city information sites, such as CitySearch and WCities [41][42]. In other words, Google Local is more like a city guide system provided on the Web, rather than a system for geographic information retrieval from the Web. MSN City Guides and Ask Jeeves Local also use data from city guide sites [36][35].

Google Earth is another geographic information service on the Web provided by Google [33]. This system provides satelite images for a wide range of the Earth's surface. It is not linked to local Web content, other than those provided by Google Local.

Yahoo! Local Maps is a map-based local Web search system provided by the major portal site, Yahoo! [34]. It is also integrated with real-time data such as traffic information. AOL Local is a service provided by America Online. It is integrated with AOL's rich content, including local events and movie information [37].

In addition to these major Web search engines and portal sites, there are also sites that provide local content as their main feature. CitySearch, mentioned above, provides well-classified regional information, through much effort by its editors [41]. Its database includes information on restaurants, hotels, and various public facilities for a number of U. S. cities. Since it requires human editing, however, it does not contain much of the valuable local information available on the Web. Switchboard provides an Internet version of the yellow pages [38]. Since its search results are not linked to Web content, it only provides a limited amount of information, such as locations and phone numbers. ShopLocal enables users to search local shop information by specifying a city name [40]. MapQuest has its own map interface and provides categorized links to location-related Web sites [39].

There are also services that enable local search of weblog sites and articles. DC Metro Blogmap and nyc bloggers provide services that link one's personal weblogs, or blogs, to metro maps for Washington, D.C. and New York City. With these services, a user can find bloggers located in specific areas in the city [44][45]. There are also services that map each blog entry to a geographic location. Uematsu et al. proposed *Ba-log*, in which users upload their blog entries by using a cellular phone equipped with a camera and a GPS extension [4]. WorldKit is a toolkit for creating map-based applications on the Web, and it has also been applied to a blog mapping service [43]. These services require manual registration and do not support automatic extraction of information from blog entries.

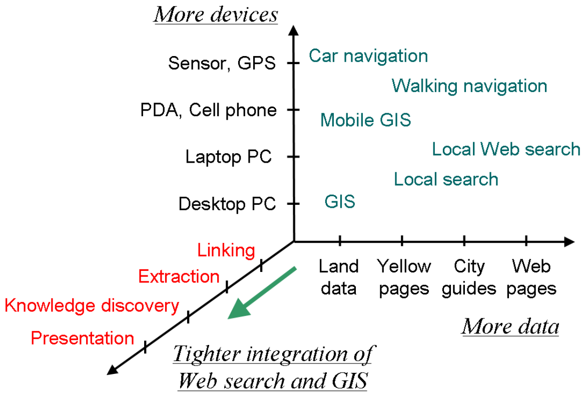


Figure 2: Directions for Web search and GIS integration

3. TIGHTER INTEGRATION OF WEB SEARCH WITH A GIS

The aim of this paper is to discuss directions for the deeper integration of Web search and a GIS. Various issues play important roles in improving the user's experience while obtaining geographic information on the Web. Figure 2 illustrates the issues involved in the integration of Web search with a GIS.

One direction of advancement is to increase the amount of data involved. Another is to increase the variety of devices used. Increasing the data quantity enables higher-resolution maps and enhances existing applications. Increasing the variety of devices enables Web search with GIS at any time, anywhere. On the other hand, tighter integration enables entirely new services and experiences for the user.

Various issues play a role in the integration process, but we have listed the four most significant ones below:

1: Linking: The simplest integration of Web search with a GIS is to link a Web page to a geographic point or region. Such integration is already implemented in many existing local Web search systems. The user can search Web content by specifying geographic location, and also find location referred by Web content.

2: Extraction: The extraction of relevant pieces of geographic information from a Web page increases user convenience by eliminating unrelated parts. It also enables further processing of the extracted data for various applications.

3: Knowledge discovery: The knowledge discovery from Web and GIS is aggregation of extracted information into knowledge on geographic space. The aggregation creates knowledge that is not available in a conventional GIS.

4: Presentation: Presenting URLs and snippets along with a map interface is the most common presentation scheme in local Web search systems. However, more useful ways of presenting Web search results with geographic information can be investigated.

Since linking is already provided by many local Web search systems, we focus on the last three issues in this paper.

4. EXTRACTION

The extraction of information from local Web search results is one of the most important issues in local Web search. At present, most local Web search systems retrieve Web pages as the search results. In many cases, however, only part of a retrieved page contains regional information. Extraction of that regional information would make the presentation to the user more efficient.

Extraction of parts of Web pages has been investigated by several groups. Sagara et al. built a system that extracts geographic terms from Web content, enabling geographic searching [5]. Woodruff and Plaunt developed the Georeferenced Information Processing SYstem (GIPSY), which parses Web documents and assigns coordinates to each page [6]. Yang and Claramunt mined user preferences regarding sightseeing spots for tourism applications [27].

Several methods can be used for extraction of geographic information:

- simple text matching;

- natural language processing of text, such as morphological analysis and structural analysis;

- ontology-based extraction, such as using dictionary data.

We next describe an application that extracts snippets of local information from Web pages.

4.1 Blog Map of Experiences

Information on what people experience at sightseeing spots is often unavailable from city guides. Such information should be of interest to tourists, however, because it would help them plan their trips. The Web is one of the best sources for gathering this type of information. We previously developed a system that extracts the region-specific experiences of people from weblogs.

The prevalence of blogs enables the accumulation of personal experiences specific to location and time. Such information has traditionally been unavailable, except indirectly through commercial city guides, local newspapers and periodicals. This kind of information is particularly valuable for potential tourists and marketing analysts interested in local trends. In addition, each blog entry has the time it was published as an attribute. This enables the extraction of the writer's experiences during a specific time period. When combined with the extraction of geographic keywords from blog entries, tourist experiences related to a specific place and time can be obtained. For example, in spring, many people visit various locations to enjoy flower blossoms. At famous sightseeing spots, a certain percentage of people enjoys local specialties. Such information is formalized as rules among place, time, action, and object. We implemented a system, the **Blog Map of Experiences**, which extracts these rules from blog entries and present them visually. The system's configuration is illustrated in Figure 3.

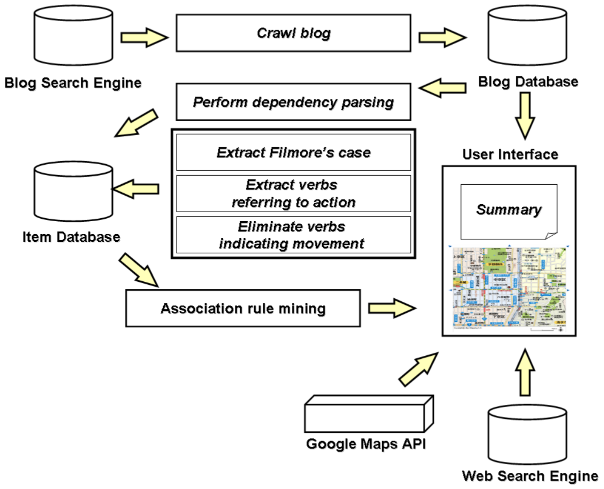


Figure 3: Blog Map of Experiences

The aim of the Blog Map of Experiences is to present tourists' real-life experiences through a map interface, in an integrated manner. Compared to conventional local information search systems, the Blog Map of Experiences has two notable features. One is that the results are based on information provided directly from a large number of blog authors. The other is that both the type of experience and the time, in addition to the location, can be specified in a search query.

Since not all of the blog entries refer to users' experiences, we restrict the search to phrases that involve actions. The following subsection describes the extraction process in detail.

4.2 System Architecture

The system first collects blog entries from blog search engines. The collected blogs are stored in a database. The system performs morphological analysis to each entry, in which sentences into words and their parts of speech are estimated.

Sentences containing these verbs indicating motion are eliminated, since their original sentences likely described motion toward a location, rather than actions at the location, which are what we want to extract and present to the users. As preprocessing for the association rule mining, which is done using the APRIORI algorithm [8], the time attributes of the entries were grouped by month, so that the resulting association rules would have higher support values.

Three types of rules can be extracted:

Type 1: [Time, Place] ⇒ [Action]
Type 2: [Time, Place, Action] ⇒ [Object]
Type 3: [Time, Place] ⇒ [Action, Object]

Other rules, such as those between actions, were not used in this system, because they do not match our goal of extracting spatially and temporally specific experiences.

The results of a preliminary experiment showed that Type-3 rules contain too much noise and could not be used without further improving the refinement methods. The precisions of these rules were around 10%.

The extraction is performed in two steps. First, Type-1 rules are extracted, and a set of typical verbs for a given place name is obtained. Second, Type-2 rules are extracted for pairs consisting of the given place name and the extracted typical verbs.

The process is described by the following pseudocode:

Code: Experience extraction

```
Define a place name set $P = p_1, ..., p_n$.
For $i = 1$ to $n$ do
        Obtain $m$ association rules $p_i \rightarrow v$,
            in decreasing order of support.
        Obtain the set of verbs $V_i = v_{i1}, ..., v_{im}$
        For $j = 1$ to $m$ do
                Obtain $k$ association rules $p_i, v_{ij} \rightarrow n$,
                    in decreasing order of support.
                Store $k$ rules.
        Done
Done
```

After the extraction, we refine the results in two ways:

Refinement 1: Identify and extract sentences referring to actions.

Refinement 2: Eliminate sentences indicating movement from one place to another.

The former refinement is applied because sentences referring to the states of the objects are not needed. The latter refinement is applied because our goal is to extract experiences at certain places, not in between.

The results are then stored in a rule database and presented to the user in various formats.

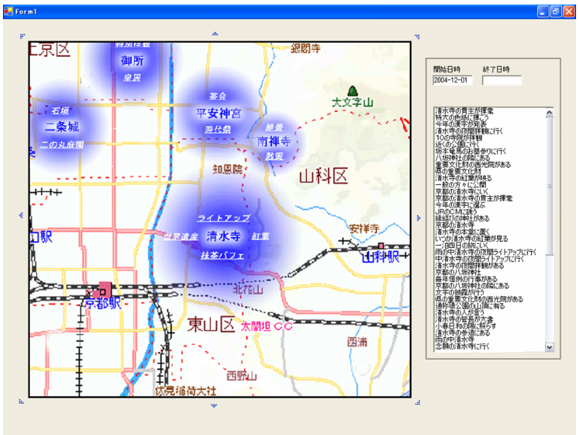
Figure 4 shows an example user interface for the Blog Map of Experiences system.

4.3 Evaluation

We evaluated our system on real data. First, we manually listed 20 popular sightseeing spots in Japan and collected 500 blog entries for each. In collecting blog entries, we used "goo blog" and "livedoor blog" [46][47], which are typical blog hosting services in Japan. The place names used in search queries are taken from digitized residential maps provided by Zenrin Co., Ltd. [48].

For morphological analysis, we used Chasen morphological analyzer [50]. We obtained a collection of verbs referring to actions from the lexical database of the Japanese Vocabulary System [51] and categorized the verbs into a tree structure, with verbs referring to actions grouped into one top-level category. We then manually selected verbs that indicate movement. For example, verbs such as "go," "come," and their synonyms are in the list.

The Type-1 rules were extracted, and the top j rules were obtained in decreasing order of *support* value. Next, the Type-2 rules were extracted for each of the verbs that were the *consequents* of the Type-1 rules, and the top ten nouns that were the *consequents* of the Type-2 rules were obtained. We calculated the average precision of the extracted pairs of verbs and nouns, i.e., the association rules. As shown in Table 1, the combination of Refinements 1 and 2 improved the precision of the rule set.



開始日時 2004-12-01
終了日時

清水寺の貫主が揮毫
特大の色紙に揮ごう
今年の漢字が発表
清水寺の夜間拝観に行く
10の寺院が拝観
近くの公園に行く
坂本竜馬のお墓参りに行く
八坂神社の隣にある
重要文化財の西光院がある
県の重要文化財
清水寺の紅葉が映る
一般の方々に公開
京都の清水寺にいく
京都の清水寺の貫主が揮毫
今年の漢字に選ぶ
JRのCMに誘う
縁結びの神社がある
京都の清水寺
清水寺の本堂に置く
いつか清水寺の紅葉が見る
一泊四日の旅にいく
雨の中清水寺の夜間ライトアップに行く
中清水寺の夜間ライトアップに行く
清水寺の夜間拝観がある
京都の八坂神社
毎年恒例の行事がある
京都の八坂神社の隣にある
文字の披露が行う
県の重要文化財の西光院がある
通称狼公園の山頂に有る
清水寺の人が言う
清水寺の管長が大書
小春日和の陽に照らす
清水寺の参道にある
雨の中清水寺
念願の清水寺に行く

Figure 4: User interface for the Blog Map of Experiences system

Top j	Size	Unrefined	Refined by 1	Refined by 1+2
3	30	0.007	0.083	0.216
5	50	0.058	0.111	0.221
10	100	0.087	0.131	0.182

Table 1: Average precision of the extracted association rules

The extracted rules can be used to recommend visitors popular activities at sightseeing spots, and also to provide analysts with means to observe local trends.

5. KNOWLEDGE DISCOVERY

Knowledge discovery is another significant topic in integrating Web search with a GIS. Aggregating extracted information into knowledge helps users learn the characteristics of a region more efficiently.

Although knowledge discovery has been widely studied in the database field, applications specific to the Web and to geographic information are limited [9]. Koperski described a method for knowledge mining in the geographic domain, yet Web content was not involved [12]. Buyukkokten et al. observed a bias in the locations of sites linked to various newspaper sites [11]. They compared the IP addresses of sites linked to the *New York Times* and the *San Francisco Chronicle* and found that the sites were more widely distributed for the *New York Times*. This is an example of regional knowledge obtained through Web mining.

There are different ways to obtain valuable geographic knowledge from the Web, as indicated below:

- by summarizing the content into an overview;

- by removing the overlaps among content;

- by identifying rules or patterns (i.e., association rules) in the content;

- by identifying the statistical tendencies in the content data.

We next describe a system that creates one of the simplest kinds of geographic knowledge, the cognitive significance of a place name. Focusing on a simple type of knowledge enables the system to obtain knowledge with a relatively small amount of data.

5.1 Place Name Ranker

In general, maps present place names that are significant within its region. Significance of a place name is cognitive information, rather than physical information. Usually, mapmakers select place names for a map manually. This is a labor intensive task, especially for a city map, where frequent updating is necessary. A selection is based on a subjective judgment, and may not match intuitions of many users. In addition, the result of a selection is qualitative, dividing place names into significant ones and non-significant ones. If the cognitive significance of place names were measured quantitatively, place names presented on a map can dynamically be selected as a map zooms.

The term *landmark* is often used to indicate a cognitively significant geographic object [20][21]. Landmarks have been studied as a part of a cognitive map, which is a mental model of geographic space that a person has in mind [17][18][19]. Such cognitive information is important in applications that involves human behavior on a geographic scale. This is indicated in that maps without names of significant geographic objects are often inconvenient for daily use.

Since the value of cognitive significance for a place name is a type of information that is not present in an ordinary GIS, we considered it an appropriate target for knowledge discovery from the Web.

We implemented a system, **Place Name Ranker**, which ranks place names by their significance based on the way they appear on Web content. The method applies conventional text mining techniques to Web pages collected by a Web crawler. Although the cognitive significance of a geographic object is a subjective measure differing from person to person, we assumed that the average figure is still useful for many applications. The following subsections define measures which are expected to represent the cognitive significance of geographic objects.

5.1.1 Document Frequency

The document frequency (DF) of a term is defined as the number of documents (Web pages) that contain the term. This measure is commonly used in text mining [14]. It is calculated as follows:

$$d(p_i) = |\{d \in D | p_i \in s \wedge s \in d\}|, \qquad (1)$$

where p_i indicates the target place name (for which the DF is calculated), D is the document set, and s is a sentence. This is one of the simplest measures of word frequency in a set of documents.

5.1.2 Regional Co-occurrence Summation

A shortcoming of the DF is that it does not examine whether a place name is used in a spatial context or not. Therefore, branches of enterprises, universities, or chain stores are often highly ranked in terms of DF. To avoid this, we need to measure the frequency with which a place name is actually used in a spatial context. We do this by calculating the regional co-occurrence summation (RS). We assume that when two neighboring place names appear in the same document, they likely are both used in a spatial context. In terms of text mining, we consider the *co-occurrence* of two neighboring place names to be an indicator of spatial context. Co-occurrence is a commonly used measure for term relationships in text mining [14, 16].

The RS is defined as the total number of co-occurrences that the target place name has with the surrounding place names. Before calculating it, we must first define the *surrounding place names*. We call this set the **physical proximity** of the target place name. One way to define this is to use a threshold distance:

$$P'(p) = \{p_i | p_i \in P_{all} \wedge \delta(p, p_i) \le R \wedge p_i \ne p\}, \qquad (2)$$

where p is the target place name, P' is the threshold-based physical proximity, P_{all} is the original set of place names, the function δ gives the distance between place names, and R indicates the threshold distance.

This simple model, however, is inappropriate if the target area contains both dense and sparse distributions of place names, because some place names will have a large number of neighboring place names and others will have only a few. As a result, this measure will have low reliability.

We thus define the physical proximity instead as *the set of n-closest place names from the target place name.* Such a set can be obtained by sorting the place names according to their distances from the target place name:

$$P(p) = \{p_i | p_j \in P_{all} \wedge \delta(p, p_j) \le \delta(p, p_{j+1}) \wedge 1 \le i \le n \wedge p_i \ne p\}. \qquad (3)$$

Next, the formula for the RS is

$$r(p_i) = \sum_{p_j \in P(p_i)} \kappa(p_i, p_j), \qquad (4)$$

where $\kappa(p_i, p_j)$ is the number of documents (Web pages) containing both p_i and p_j. In other words, $\kappa(p_i, p_j)$ is the number of co-occurrences between p_i and p_j, in terms of documents.

The use of the RS reduces the effect of ambiguity in place names. Suppose that place name a indicates two different coordinates, $\boldsymbol{x_a}$ and $\boldsymbol{x_{a'}}$, while place name b indicates coordinate $\boldsymbol{x_b}$. Suppose also that the distances between the three coordinates follow the order $|\boldsymbol{x_a} - \boldsymbol{x_b}| < |\boldsymbol{x_{a'}} - \boldsymbol{x_b}|$. If a and b co-occur in document A, a in document A likely refers to coordinate $\boldsymbol{x_a}$, rather than to $\boldsymbol{x_{a'}}$. Because the DF does not account for such ambiguity, we expect the RS to perform better than the DF in extracting significance in a spatially restricted sense.

Although various distances can be defined (i.e., the network metric distance and the time distance), we use the Euclidean distance between coordinates, since the data necessary for calculating the other distances are not as easily obtained for many regions.

5.2 System Architecture

Based on these two measures, we implemented our Place Name Ranker system. The main components are shown in Figure 5. The basic data used by the system are (1) place names and their coordinates from a GIS and (2) text content from the Web.

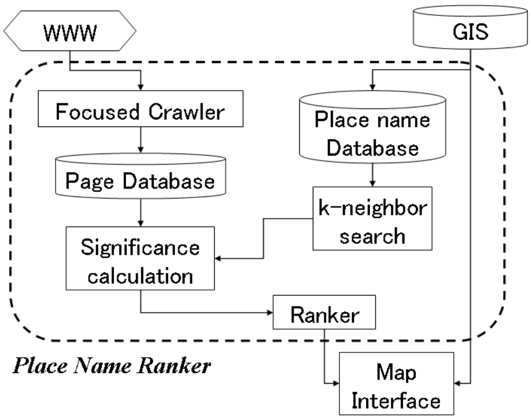


Figure 5: Architecture of Place Name Ranker

A focused crawler is a special type of Web crawler that collects pages that only fulfill certain conditions. In other words, links are traced only when a page satisfies these conditions. The conditions can include the existence of certain keywords or high similarity to given example pages. Focused crawlers have been reported to have higher efficiency in retrieving Web pages for certain topics [13]. The underlying assumption is that pages concerning certain topics are more likely to be linked to pages discussing the same topic.

Because our objective is to extract significant place names for a certain target region, we use a focused crawler for efficient retrieval, and we set one condition-the page must contain at least one place name from within the region of interest. Due to the ambiguity of place names, not all of the retrieved pages will discuss the target region. Nevertheless, the ratio of related pages is higher than with a conventional crawler.

After measuring the values for cognitive significance, the map interface presents the place names with the top k measured values.

5.3 Evaluation

We tested this system by determining how well it could assign values to place names. To evaluate the appropriateness of the proposed measures, DF and RS, we performed an experiment using the data indicated below.

Participants: 36 residents of the target region, plus 14 people from outside the region.

Responses: Each participant was asked to select the 20 most significant place names in the target region. The result contained 275 distinct place names.

GIS data: The place name data were taken from a regular GIS, a digitized residential map for Kyoto, Japan, provided by Zenrin Co., Ltd. [48]. The map data were divided into layers, including a "significant objects" layer containing 7109 place names. Although we could assume objects in this layer to be significant, their levels of significance varied. Famous temples and ordinary elementary schools were all included in the same layer, and no quantitative value of significance was assigned to the objects. Our goal in this experiment was to see if the system could assign a significance value to each of these names.

Web documents: As a document set, we used 157,297 Web pages collected by the focused crawler. Only the text parts were used in the information extraction. The total data size was 2.45 GB. To calculate the DF, we used the Namazu full-text search engine [49]. Prior to the search, we added place names taken from the GIS to the Namazu indexed word list. The RS was calculated in the same manner.

We used precision and recall to evaluate the two proposed measures. The definitions for the precision and recall are as follows:

$$\text{Precision} = \frac{\text{No. of Retrieved Correct Objects}}{\text{No. of Retrieved Objects}}, \quad (5)$$

$$\text{Recall} = \frac{\text{No. of Retrieved Correct Objects}}{\text{No. of Correct Objects in Population}}. \quad (6)$$

We sorted the place names taken from the GIS in decreasing order of the two measures. Rank position k indicates that the precision and recall were calculated for the top k entries in the ranking. If k was small, the set was likely to have high precision and low recall, while if k was large, the opposite situation was likely. The precision and recall are functions of k. A precision-recall curve (P-R curve) is commonly used to visualize a series of precision and recall pairs obtained by varying k [16].

We then sorted the place names from the GIS in decreasing order of the calculated DF and RS values. The top k place names were selected, and their precision and recall values were calculated with respect to a set of significant place names judged manually.

We graphed the points where the recall increased in order to make the P-R curve smooth. By renumbering the extracted pairs, we obtained a series of P-R pairs as a function of a new parameter, j. We then averaged the P-R pairs collected from different participants for each j and obtained an averaged P-R curve. This curve was a function of j. This method is called *averaging by micro-evaluation* [16]. The value of k ranged from 1 to 7109 (= the number of significant place names in the GIS), j ranged from 1 to 20 (= the number of "correct answers" given by each participant), and the number of P-R pairs (to be averaged) was 50 (= the number of participants).

The averaged P-R curves for the DF and the RS are shown in Figure 6, and the precisions at different rank positions are listed in Table 2. These results indicate that that RS performed relatively well as a measure for evaluating the cognitive significance of place names, especially when compared to the DF. A more detailed discussion is given elsewhere [31].

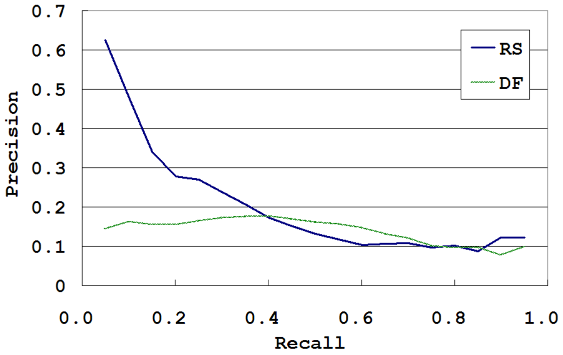


Figure 6: P-R curves for the DF and the RS

Rank pos.	DF	RS
5	0.016	0.368
10	0.140	0.212
15	0.096	0.259
20	0.106	0.239

Table 2: Precisions at different rank positions

The results show that by selecting an appropriate measure, geographic knowledge can be effectively extracted from the Web.

6. PRESENTATION

A map interface is not the only way to present the result of local Web search. There are various other means of presentation, in different styles and media.

A number of methods have been proposed as new presentation schemes for geographic information, especially for presentation on mobile devices Since mobile devices have limited display capacity, the presentation method is of great importance. Kim et al. implemented a mobile tour-guide system that automatically finds Web services and presents the integrated results on a map interface [23]. Gardiner and Carswell developed a system that processes directional queries with respect to the user's view [24]. Their Cultural Heritage Interface enables the user to obtain information about locations by specifying directions with respect to himself or herself. Huang et al. implemented a system that integrates a GIS, virtual reality (VR), and the Internet through the Virtual Reality Modeling Language (VRML) [25]. The integration discussed in their paper covers only VRML data and does not involve other types of media on the Web, such as text.

The presentation schemes for the Web with GIS, however, can proceed in several different directions:

1. unrestricted by maps (the map interface is not the only way to present geographic content);

2. unrestricted by (conventional) Web browsers (the Web browser is not the only way to present local Web search results);

3. unrestricted by explicitly specifying queries (proactive presentation).

The last item indicates that most of the conventional digitized map applications are built on the query-and-response model. The map interface waits for a user query and presents the result in response. That is, these are query-driven systems.

In the following subsections, we describe an example of a new mode of presentation for the Web search integrated with a GIS. This system uses neither a map nor a Web browser, and it is also proactive presentation.

6.1 Web Car Radio

The **Web Car Radio** enables users to listen to Web content as they drive a vehicle. The location of the vehicle is constantly tracked using GPS. As with listening to a car radio, a user can select various channels, each focusing on various topics of interest, such as restaurants, events, and sightseeing spots. The concept of this system is illustrated in Figure 7.

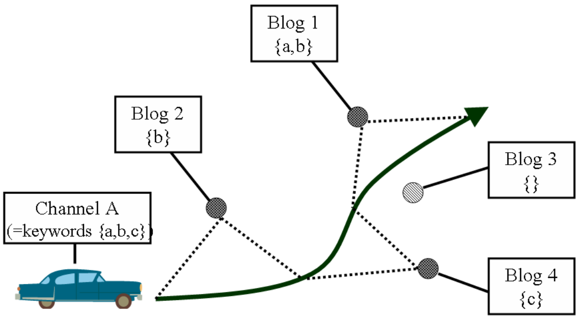


Figure 7: Web Car Radio

The Web Car Radio has a user interface that uses a vehicle's motion as input. The user obtains Web content without explicitly specifying queries. In return, the system presents Web content through sound by using speech synthesis. The system is thus based on the concept of proactive or query-free browsing [26].

6.2 System Architecture

Web Car Radio extracts regional information from the Web and presents it to the driver and his fellow passengers by speech synthesis.

The primary source of information used is personal diary sites on the Web. Since diaries are generally expressed in a conversational style, we considered them appropriate for presentation as radio programs. Many of Web diary sites today are provided as Weblogs (blogs), and data are available in the RSS style. Now that there are many non-commercial RSS search engines, the content for Web Car Radio can easily be collected [46][47].

In the process, the system first map blog entries to geographic locations. By using various location identification techniques, the sytem maps the addresses and place names contained in a blog entry[1][7]. Some blog entries now have spatial tags as well, making them easier to use, although the tags may not contain all of the locations to which a blog entry refers.

The location of the vehicle is obtained by GPS. Geographic data in the GIS provide the set of addresses and well-known place names located close to the vehicle's location. Next, the system performs a Web search and obtain Web pages. It extracts phrases that contain addresses or place names, as well as their surrounding phrases. The extracted phrases are then provided through audio by speech synthesis. The audio output is suitable for the driver, since he or she is unable to watch a screen.

Selecting a channel on Web Car Radio corresponds to selecting keywords for a Web search. Each channel has a set of keywords representing certain topics. The system uses these keywords to filter the retrieved search results. The relevance between the document and the selected channel is calculated by the cosine measure [15]. The system also selects Web content that contains place names, and rank them in the order of proximity to the vehicle's present location. For example, if the user sets "restaurants" as the topic, the Web content concerning local restaurants is searched and presented by speech synthesis.

Entries are ordered by the measure $m(x_c, p, D, \mathbf{v_k})$, which is calculated as:

$$m(x_c, p, D, \mathbf{v_k}) = -k_a d(x_c, x(p)) + k_b S^{(C)}(\mathbf{v_k}, \mathbf{v_D}), \quad (7)$$

where d is the distance function, x_c is the current vehicle location, $x(p)$ is the location of the place name p contained in the document D, $\mathbf{v_k}$ is the selected channel's keyword vector, $\mathbf{v_D}$ is the document vector for document D, $S^{(C)}$ is the cosine similarity measure, and k_a and k_b are arbitrary positive coefficients, optimized through learning.

Once an entry is selected, it is presented to its end, since the discontinuity in speech would confuse the user. While the vehicle is moving, the measurement is repeatedly calculated after a set time interval t passes.

Figure 8 illustrates the architecture for the Web Car Radio system. The input comes from the GPS and the channel selector. The output is synthesized speech from the vehicle's audio devices. Topic ontology database is used to construct the set of keywords for each channel.

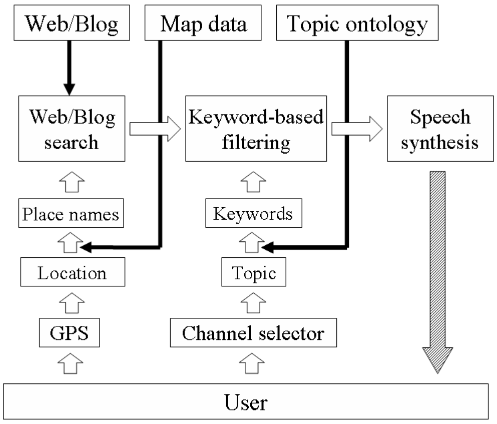


Figure 8: Architecture of Web Car Radio

There has been much research on retrieving neighboring objects for a vehicle [28][29][30]. Our system, however, has the characteristics that the search domain is the Web and the content is presented by audio, requiring much less effort from the user.

7. CONCLUSION

We have discussed the issues involved in integrating Web search with a geographic information system (GIS). Specifically, the issues of extraction, knowledge discovery, and presentation were covered.

We have implemented a system that obtains characteristic experiences from weblog entries. We showed that characteristic experiences can be extracted for different locations. For knowledge discovery, we measure the cognitive significance of geographic objects according to content gathered from the Web. We also discussed possible presentation methods, including an application example that is not restricted by the limitations of a map interface or Web browser.

The future directions presented here are neither exhaustive nor conclusive. Tighter integration will lead to various applications that are not possible in either conventional Web search or in a GIS. The integration of Web search with a GIS is thus very promising.

8. ACKNOWLEDGMENTS

This work was supported in part by the Japanese Ministry of Education, Culture, Sports, Science and Technology under a Grant-in-Aid for Software Technologies for Search and Integration across Heterogeneous-Media Archives, and under a Special Research Area Grant-In-Aid For Scientific Research (2) for the year 2005, as part of a project titled Research for New Search Service Methods Based on the Web's Semantic Structure (Project No. 16016247; Representative, Katsumi Tanaka); and by the Informatics Research Center for Development of Knowledge Society Infrastructure (COE Program of the Japanese Ministry of Education, Culture, Sports, Science and Technology).

9. REFERENCES

[1] McCurley, K. S. Geospatial mapping and navigation of the Web, in Proceedings of the 10th International World Wide Web Conference (WWW10), 221-229, Hong Kong, China, 2001.

[2] Larson, R. R. Geographic information retrieval and spatial browsing, Smith, L. and Gluck, M. (eds.), GIS and Libraries: Patrons, Maps and Spatial Information, 81-124, University of Illinois, 1996.

[3] Plewe, B. GIS-Online: Information Retrieval, Mapping, and the Internet, OnWord Press, 1997.

[4] Uematsu, D., Numa, K., Tokunaga, T., Ohmukai, I. and Takeda, H. Ba-log: a proposal for the use of locational information in blog environment, in Proceedings of the 6th Web and Ontology Workshop, Japanese Society for Artificial Intelligence, 2004.

[5] Sagara, T., Arikawa, M. and Sakauchi, M. Spatial information extraction system using geo-reference information. Information Processing Society of Japan Journal: Database, Vol. 41, No.SIG6(TOD7), 69-80, 2000.

[6] Woodruff, A. G. and Plaunt, C. GIPSY: Automated geographic indexing of text documents. Journal of the American Society for Information Science, 45, 9, 645-655, 1994.

[7] Kanada, Y. A method of geographical name extraction from Japanese text for thematic geographical search, in Proceedings of the 8th International Conference on Information and Knowledge Management, 46-54, Kansas City, Missouri, 1999.

[8] Agrawal, R. and Srikant, R. Fast algorithms for mining association rules in large databases, in Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94), 487-499, 1994.

[9] Piatestsky-Shapiro, G. and Frawley, W. J. (eds.). Knowledge Discovery in Databases, MIT Press, Boston, Massachusetts, 1991.

[10] Lim, E. P., Goh, D., Liu, Z., Ng, W. K., Khoo C. and Higgins, S. E. G-Portal: A map-based digital library for distributed geospatial and georeferenced resources, in Proceedings of the 2nd ACM+IEEE Joint Conference on Digital Libraries (JCDL 2002), 351-358, Portland, Oregon, 2002.

[11] Buyukkokten, O., Cho, J., Garcia-Molina, H., Gravano, L. and Shivakumar, N., Exploiting geographical location information of Web pages, in Proceedings of the ACM SIGMOD Workshop on the Web and Databases (WebDB'99), Philadelphia, Pennsylvania, 1999.

[12] Koperski, K., Han, J. and Adhikary, J. Mining knowledge in geographical data. Communications of the ACM, 26, 1, 65-74, 1998.

[13] Chakrabarti, S., van den Berg, M. and Dom, B. Focused crawling: a new approach to topic-specific Web resource discovery, in Proceedings of the 8th International World Wide Web Conference (WWW8), Toronto, Canada, 1999.

[14] Salton, G. Automatic Information Organization and Retrieval, McGraw-Hill Inc., 1968.

[15] Salton G. and Buckley, C. Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24, 5, 513-523, 1988.

[16] van Rijsbergen, C. J. Information Retrieval - Second Edition, Butterworth & Co Publishers Ltd, 1979.

[17] Lynch, K. The Image of the City, MIT Press, 1960.

[18] Neisser, U. Cognition and Reality: Principles and Implications of Cognitive Psychology, W. H. Freeman and Company, 1976.

[19] Couclelis, H., Golledge, R., Gale, N. and Tobler, W. Exploring the anchor-point hypothesis of spatial cognition. Journal of Environmental Psychology, 7, 2, 99-122, 1987.

[20] Raubal, M. and Winter, S. Enriching wayfinding instructions with local landmarks. Geographic Information Science, Lecture Notes in Computer Science 2478, 243-259, Springer-Verlag, 2003.

[21] Elias, B. Extracting landmarks with data mining methods. Spatial Information Theory: Foundations of Geographic Information Science, Lecture Notes in Computer Science 2825, 375-389, Springer-Verlag, 2003.

[22] Avesani, P., Cova, M., Hayes, C. and Massa, P. (eds.). Proceedings of the WWW2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, Chiba, Japan, 2005 (entire issue).

[23] Kim, J. W., Kim, C. S., Gautam, A. and Lee, Y. Location-based tour guide system using mobile GIS and Web crawling. Web and Wireless Geographical Information Systems, Lecture Notes in Computer Science 3428, 51-63, Springer-Verlag, 2005.

[24] Gardiner, K. and Carswell, J. D. Viewer-based directional querying for mobile applications, in Proceedings of the 3rd International Workshop on Web and Wireless Geographical Information Systems (W2GIS 2003), 73-83, Rome, Italy, 2003.

[25] Huang, B., Jiang, B. and Lin, H. An integration of GIS, virtual reality and the Internet for spatial data exploration. International Journal of Geographical Information Science, 15, 5, 439-456, 2001.

[26] Nadamoto, A., Kondo, H. and Tanaka, K. WebCarousel: Restructuring Web search results for passive viewing in mobile environments, in Proceedings of the 7th International Conference on Database Systems for Advanced Applications (DASFAA 2001), Hong Kong, China, 2001.

[27] Yang, Y. and Claramunt, C. A Hybrid Approach for Spatial Web Personalization. Web and Wireless Geographical Information Systems, Lecture Notes in Computer Science 3428, 206-221, Springer-Verlag, 2005.

[28] Zhang, J., Zhu, M., Papadias, D., Tao, Y. and Lee, D. L. Location-based spatial queries, in Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, 443-454, San Diego, California, 2003.

[29] Mokbel, M. F., Xiong, X. and Aref, W. G. SINA: Scalable incremental processing of continuous queries in spatio-temporal databases, in Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, 623-634, Paris, France, 2004.

[30] Ishikawa Y., Tsukamoto, Y. and Kitagawa, H. Implementation and evaluation of an adaptive neighborhood information retrieval system for mobile users, in Proceedings of the 3rd International Workshop on Web and Wireless Geographical Information Systems (W2GIS 2003), 17-26, Rome, Italy, 2003.

[31] Tezuka, T. and Tanaka, K. Landmark extraction: a Web mining approach. Spatial Information Theory, Lecture Notes in Computer Science 3693, 379-396, Springer-Verlag, 2005.

[32] Google Local, http://local.google.com/

[33] Google Earth, http://earth.google.com/

[34] Yahoo! Local Maps, http://maps.yahoo.com/

[35] Ask Jeeves Local, http://local.ask.com/

[36] MSN City Guides, http://local.msn.com/

[37] AOL Local Search, http://localsearch.aol.com/

[38] Switchboard, http://www.switchboard.com/

[39] MapQuest, http://www.mapquest.com/

[40] ShopLocal, http://www.shoplocal.com/

[41] CitySearch, http://www.citysearch.com/

[42] WCities, http://www.wcities.com

[43] WorldKit, http://www.brainoff.com/worldkit/index.php

[44] DC Metro Blogmap, http://www.reenhead.com/map/metroblogmap.html

[45] nyc bloggers, http://www.nycbloggers.com/

[46] goo blog, http://blog.goo.ne.jp

[47] livedoor blog, http://blog.livedoor.com/

[48] Zenrin Co., Ltd., http://www.zenrin.co.jp/

[49] Namazu: a Full-Text Search Engine, http://www.namazu.org/index.html.en

[50] Chasen, http://chasen.aist-nara.ac.jp/index.html

[51] Japanese Vocabulary System, http://www.ntt-tec.jp/technology/C404.html