Landmark Extraction: A Web Mining Approach

Taro Tezuka and Katsumi Tanaka

Graduate School of Informatics,
Kyoto University
{tezuka, tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract. Landmarks play crucial roles in human geographic knowledge. There has been much work focusing on the extraction of landmarks from geographic information systems (GIS) or 3D city models. The extraction of landmarks from digital documents, however, has not been fully explored. The World Wide Web provides a rich source of region related information based on our understanding of geographic space. Web mining enables a new mean of extracting landmarks, differently from conventional vision oriented methods. Our approach is based on how geographic objects are *expressed* by humans, instead of how they are *observed*. We extend existing methods of text mining so that spatial context is considered. The results of the experiments showed that adopting spatial context into text mining improves the precision of extracting landmarks from web documents.

1 Introduction

Cognitive geography has interested many researchers from various fields, including civil engineering, geography, cognitive science, sociology, and marketing [7,12,35]. Researchers are interested in this subject because human spatial behavior is often based on a cognitive image of space, rather than on the actual physical structure. People act according to how they understand their environment. A pioneering work in this field is that of Lynch, a civil engineer, who uncovered basic elements of a city image from questionnaires collected from local residents [15].

From a theoretical viewpoint, Egenhofer and Mark described the characteristics of *naive geography*, a system greatly different from physical geography [8]. Mark and Frank discussed cognitive geographic space based on the recent achievements of cognitive linguistics [16].

Cognitive geography is increasingly important in the new applications of geographic information systems (GIS). Until recently, GIS has been a specialized tool for trained users such as scientists and city planners. Now more and more people use GIS for daily activities, including car navigation, pedestrian navigation, and a map service over the Internet. For these new applications, cognitive information plays an important role in making the map easier to understand for untrained users.

Conventional work on uncovering of cognitive geography, however, was mainly based on questionnaires. Such an approach is not directly applicable for practical purposes in landmark extraction, because collection and analysis of questionnaires are often cumbersome, labor-intensive tasks. In this paper, we determine the capability of extracting such information from digital documents collected from the World Wide

A.G. Cohn and D.M. Mark (Eds.): COSIT 2005, LNCS 3693, pp. 379–396, 2005.

Web. The Web today contains a tremendous amount of region related document, and it is continuously expanding.

Although cognitive geography consists of a wide variety of elements, we limited our target to landmarks. The Oxford English Dictionary defines a landmark as follows [24]:

Landmark: An object in the landscape, which, by its conspicuousness serves as a guide in the direction of one's course (orig. and esp. as a guide to sailors in navigation); hence, any conspicuous object which characterizes a neighborhood or district.

Some of the important characteristics of landmarks are as listed below.

- Cognitively significant.
- Visually salient.
- Used in navigation tasks.
- Used for determining the direction.
- Has a specific location and is often abstracted as a point.

In this paper, we consider a landmark to be a cognitively significant geographic object that is geometrically categorized as a point. This is an abstraction. Some landmarks may have relatively large spatial extensions, for example Champs Elysee in Paris or the River Thames in London. However, we consider them as a point too. In a large scale Champs Elysee or the River Thames must be considered as regions, yet in a smaller scale, they can be considered as points.

The importance of landmarks in geographic cognition has been discussed in many literatures. Tom and Denis compared street and landmark information in giving directions, and concluded that landmark oriented directions are more effective in many cases [34]. Michon and Denis discussed in what situation landmarks become effective means of giving directions [19]. However, the importance of landmarks is not limited to the way findings.

Neisser pointed out that cognitive maps are useful tools for *memorizing* geographic knowledge [22]. Indeed, much of human geographic knowledge is said to be stored with respect to landmarks and other cognitively significant geographic objects, rather than by coordinates [15]. This is quite different from conventional GIS data structures. Figure 1 shows two models for storing geographic data. The one on the left is a coordinates-oriented model, on which most conventional GISs are based on. The one on the right is a landmark-oriented model, which we assume corresponds to most of human geographic knowledge. In our model, landmarks are linked to each other by *spatial relationships*. These relationships include topological ones such as *inside of*, geometrical ones such as *close to*, and directional ones such as *to the north of*. The location of each landmark is thus determined in relation to other landmarks. Landmarks have *neighborhoods*, which are areas considered to be close enough from the landmark. The criteria for the closeness vary among observers, yet the distance in physical space is one common factor. Locations of many insignificant geographic objects are memorized using the neighborhoods of the landmarks. Such hierarchical structure in cognitive geography has been discussed for example in anchor-point theory by Counclelis et al [5].

We propose an automated landmark extraction method based on the usage of landmarks in digital documents. We collected documents from the World Wide Web and evaluated different measurements that could be used for the landmark extraction.

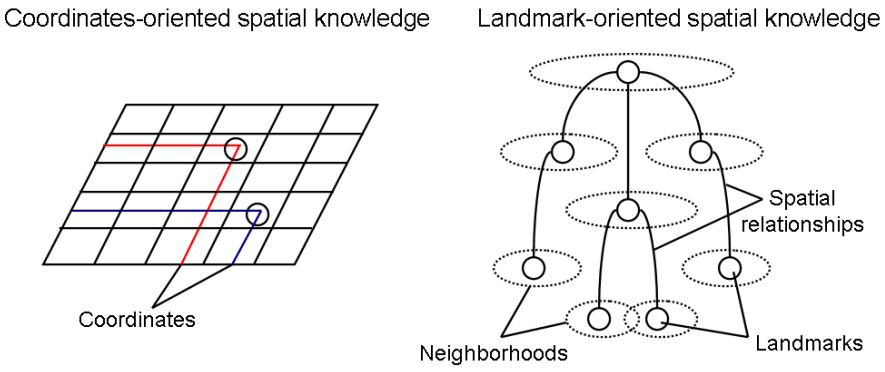

Fig. 1. Coordinates- and landmark-oriented models for geographic space

2 Related Work

The extraction of landmarks from spatial data has interested many researchers (Figure 2). Burnett and May asked subjects to write descriptions of paths to (1) their familiar locations and (2) an unfamiliar location along a path shown by a video. From this collection of descriptions, they manually extracted landmarks and also asked the subjects why they chose these objects as landmarks [1]. Raubal and Winter developed a combined method employing a 2-D GIS, photographs of road intersections, and a prominent architecture database. Indicators such as the size of a building's facade, colors, architectural importance, and shape deviation from a rectangle were used to determine the overall significance of landmarks. Statistical tests were applied to judge whether the difference from the environment was large enough [25]. Brenner and Elias used cadastral maps and airborne laser scanning data to obtain layouts and height information for various buildings. They then applied data mining techniques, such as ID3 and clustering, and obtained visually significant objects and the sizes of the areas in which these landmark can be seen [3,9]. Koiso et al. extracted landmarks according to occupancy of the visual field and categorical differences from the environment [13]. This approach is based on a hypothesis that an object that is visually significant and that belongs to a different category from the surrounding environment is more likely to be a landmark. Moon et al., in dealing with robot navigation, pointed out that the vertical lines of objects can be used as a good indicator of landmarks, even though they are much smaller in scale than the typical geographic scale. These landmarks, incidentally, are used by a robot to navigate their way through a workspace [21]. Finally, Sorrows and Hirtle provided a good survey on what is necessary for a geographic object to become a landmark [30]. Their list of landmark characteristics included *singularity*, *prominence*, *accessibility*, *content*, and *prototypicality*.

There has been extensive research on the extraction of region related information from the Web. Most of the research, however, focused on providing users with a set of web pages related to certain area or theme [2,28,17,14]. For example, Georeferenced Information Processing SYstem (GIPSY) [37] is a system similar to ours in that it parses through documents and retrieves place names and their characteristics. MetaCarta is a

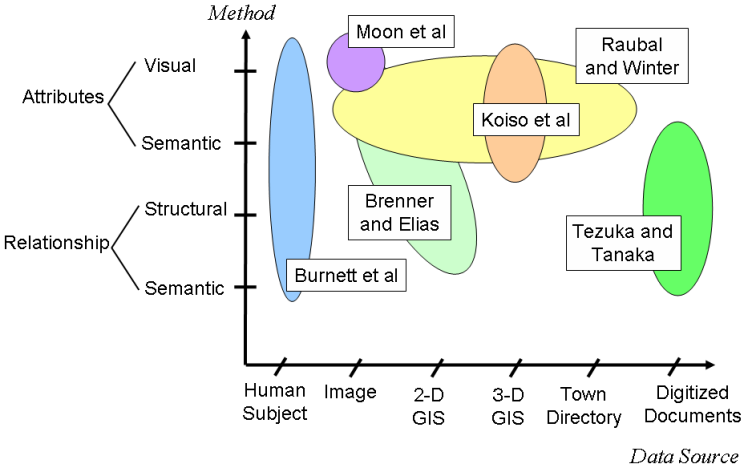

Fig. 2. Methods for landmark extraction

commercial geographic information retrieval system where documents from the Web and other sources can be searched based on the place names contain [18]. However, the aim of these systems was to assign geographical coordinates to documents. Our focus is on extracting information and obtaining new knowledge on cognitive geographic space.

Text mining research has dealt with extracting significant *terms* from documents. The text mining methods extract terms that are significant in the general sense [29,31]. In this paper, we discuss modifying existing text mining methods to include spatial context, in order to obtain spatially significant geographic objects from a very large corpus such as the Web.

3 Characteristics of Our Approach

Conventional methods for landmark extraction have focused mainly on modeling how landmarks are perceived by people. The basic idea was to model human perception and to implement a system that imitates the process of landmark cognition. Figure 3 shows the schema for this approach. The problem with this approach is that it is considering only a partial structure of landmark-human interaction. Much research has pointed out that the visual significance is not the only factor that determines landmarks. For example, despite their visual significance, skyscrapers do not always become landmarks. Another problem is that the process of modeling always encompasses selecting a limited number of attributes and ignoring the others.

In this paper, we focus on the usages of landmarks. We propose a model that landmarks are objects that are not only visually significant but also those that are frequently *used* by people. Landmarks have a variety of uses. First, they are used in organizing geographic knowledge, as described in the previous section. Second, they are used for finding one's way. Third, they are used for communication. People discuss certain locations by referring to nearby landmarks. Expressions such as *near A* are commonly used

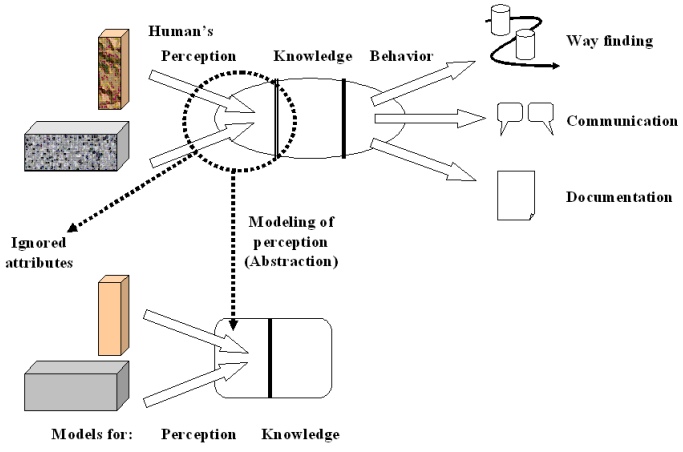

Fig. 3. Perception-based extraction of landmarks

Table 1. Uses of landmarks

Organizing concept	Organization of spatial knowledge
Way finding	Navigation to a destination
Communication	Description of regional knowledge to others
Symbol	Symbol of a region, city, or district

if *A* is a significant landmark. Lastly, landmarks are used as a symbol of either a city or a small district within a city. Table 1 lists some of the prominent uses of landmarks.

The emergence of a landmark is rather a circulatory process, in which perception is influenced by one's actions, as indicated in Figure 4. Evidence from cognitive science suggests that people are more likely to recognize objects that they expect [27]. Applying this to geographic level, objects are more likely to be recognized if the observer already knows them. Thus the objective properties such as visual significance are not the only factors that affect the emergence of landmarks. Familiarity with the object, the behaviors involved, and communication with other people play important roles. This is a *perceptual cycle*, as described by Neisser, where the significance of geographic objects increases as they are repeatedly used [22].

In extracting landmarks, not only their visual significance should be considered, but also their interaction with humans. Figure 5 illustrates the characteristics of our approach.

Because it is still difficult to trace all of human actions related to landmarks (barring drastic advancement in measurement technologies), we focus only the documentation activities of the landmarks.

Today, the World Wide Web provides a rich source of region related document. Our approach uses the Web for extracting significant landmarks to overcome the limit of perception-oriented landmark extraction methods. While most existing methods in landmark extraction are aimed at estimating how humans *observe* each geographic object, our method focuses on how people *express* landmarks.

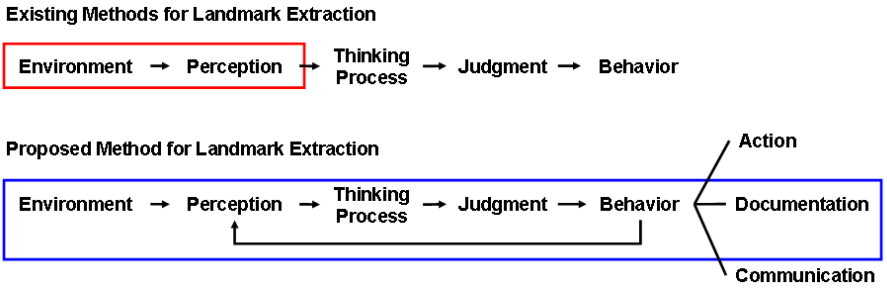

Fig. 4. Perception cycle in landmark emergence

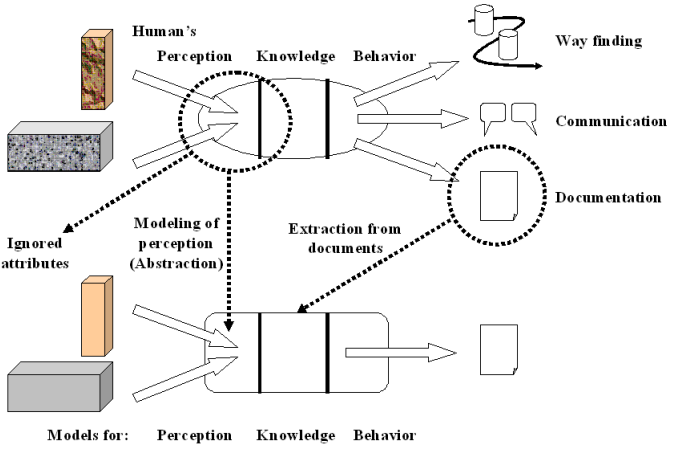

Fig. 5. Document-enhanced extraction of landmarks

In this paper, we employ text mining methods to extract significant landmarks from web documents. We have extended conventional measures in text mining so that the **spatial context** is considered.

Figure 6 shows a model that relates web, cognitive, and physical space. It asserts that the web content does not directly match the physical world. Rather, the web content is based on geographic knowledge owned by the creators of the contents.

4 Measures for the Cognitive Significance of Geographic Objects

In this section we describe the measures that we employ in extracting landmarks from web documents. First we discuss some of the general aspects of the measures, and then we describe each measure in more detail.

4.1 Landmark as a Relative Concept

Being a landmark is not a definite attribute. Whether a geographic object is observed as a landmark depends largely on the knowledge of the observer, his/her purpose, the

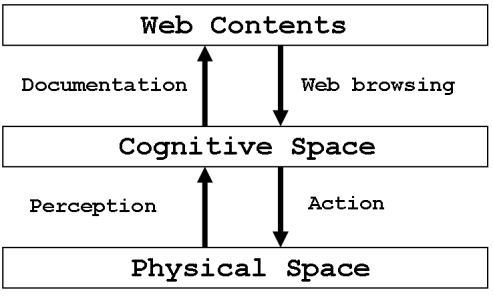

Fig. 6. Three-layer model: web content, cognitive space, and physical space

means of transportation that he/she employs, or even the time of the day. More importantly, it differs by the scale of the area being considered.

When discussing a small area, even a traffic sign or a grocery store sometimes acts as landmarks. On the other hand, there are highly significant landmarks that symbolize a city or even a state. In general, landmarks are objects that are significant *in respect to* the environment.

Thus, instead of obtaining static and absolute sets of landmarks, it is preferable to assign each geographic object a value that indicates its level of cognitive significance. In this way, whenever a range query is given, the system can return the objects that are most significant in the range query.

Before going into the discussion on the measures, however, we first compare two types of cognitive significance, general and spatial.

4.2 General and Spatial Significance

In text mining, the document frequency (DF) has been widely used as a measure for a word's commonness. The DF represents the number of documents in which a term appears. When applied to proper nouns, the popularity of an object can be estimated using the DF. However, this measure is insufficient in terms of measuring an object's significance in spatial context, for several reasons, since 1) well-known geographic objects are not necessary significant under the direct observation in the physical space, 2) the names of geographic objects often have ambiguities, where two different objects have a common name, 3) a single geographic objects may have more than one name, such as the official name and popular alternative names. In this paper, we introduce **spatial significance** in addition to **general significance** to cope with such problems. The difference between the two is as follows.

General significance: There is a class of geographic objects that are well-known in general sense but not as much in spatial sense. For example, although branches of enterprises and universities have specific locations in space, they are well-known not for its spatial properties but rather for other properties.

Spatial significance: Landmarks, nodes of traffic, significant paths, and characteristic regions are all significant in spatial sense. People know their locations, and

they play significant roles in people's spatial knowledge. These are objects that local residents can easily locate on a map.

The examples below indicate a difference between general and spatial use of a term "McDonald's".

- Our store is next to a McDonald's.
- McDonald's says its beef is safe.

The former sentence indicates that the McDonald's is sometimes used as a spatially significant object in the area, while the latter sentence doesn't have such indication.

The aim of our research was to extract geographic objects that have high spatial significance by introducing spatial context into web mining techniques. We created four measures that consider spatial context, and compared them with a measure that does not consider spatial context.

From this point on, we use the term *place name* to indicate the names of geographic objects as they are expressed in web documents. In our method, the extraction of place names itself from sentences is not involved. Instead, our goal is to measure the level of significance for each given place name, when a set of place names is available from GIS.

Text mining methods can roughly be classified into two groups, **statistical** and **linguistic** [20]. The former takes a document as a set of terms, while the latter also uses the structures of the sentences. In our measures, the first three are statistical, and the latter two are linguistic. The characteristics of the five measures are listed in Table 2.

Table 2. Characteristics of the measures

Measure	Spatial context	Category
Document frequency (DF)	No	Statistical
Regional co-occurrence summation (RS)	Yes	Statistical
Regional co-occurrence variation (RV)	Yes	Statistical
Spatial sentence frequency (SF)	Yes	Linguistic
Case frequency (CF)	Yes	Linguistic

4.3 Definitions of Measures

In this subsection, we describe five measures that are expected to reflect significance of place names as landmarks. For each measure, we give an underlying hypothesis that advocates that the measure is suitable for obtaining landmark significance.

1) Document Frequency (DF)

Underlying Hypothesis: Landmarks are frequently mentioned in web documents.

The document frequency (DF), as described in the previous subsection, is defined for each term as the number of documents (web pages) in which the term appears. This measure is commonly used in text mining [29]. The formula for the DF is as follows:

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

Here, p_i indicates the target place name (for which the DF is calculated), D is the document set, and s is a sentence.

2) Regional Co-occurrence Summation (RS)

Underlying Hypothesis: Landmarks are frequently mentioned in web documents together with place names in its surrounding.

A problem with the DF is that it does not examine whether a place name is used in a spatial context or not. Therefore, it is often the case that branches of enterprises, universities, or chain stores come highly ranked by the DF. In order to avoid such inappropriateness, we want to measure the frequency that a place name is actually used in spatial context. In calculating a regional co-occurrence summation (RS), we assume that when two neighboring place names appear in a same document, it is likely that these two place names were both used in spatial context. In terms of text mining, we consider a *co-occurrence* of two neighboring place names as an indicator of spatial context. Co-occurrence is a commonly used measure for term relationships in text mining [29,26].

The RS is defined as the total number of co-occurrences that the target place name has with its surrounding place names.

Before calculating the RS, we must first define the *surrounding place names* of the target place name. We call this set the **physical proximity** of the target place name. One way to define this is to use a threshold distance. The formula is as follows.

$$P'(p) = \{p_i | p_i \in P_{all} \wedge \delta(p, p_i) \leq R \wedge p_i \neq p\}$$

Here, p is the target place name, and 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 model becomes inappropriate if the target area contains both dense and sparse distributions of place names. In such case, there will be place names with a large number of neighboring place names, while some other place names have only few neighbors. As a result, the measure will have a low reliability.

Instead, we define the physical proximity 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 distance from the target place name. The formula for this definition is as follows.

$$P(p) = \{p_i | p_j \in P_{all} \wedge \delta(p, p_j) \leq \delta(p, p_{j+1}) \wedge 1 \leq i \leq n \wedge p_i \neq p\}$$

The formula for the RS, denoted by $r(p_i)$, is as follows.

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

Here, $\kappa(p_i, p_j)$ 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 the ambiguities in place names. Suppose that a place name a indicates two different coordinates, $\boldsymbol{x_a}$ and $\boldsymbol{x_{a'}}$, while place name b indicates coordinates $\boldsymbol{x_b}$. Suppose also that the distances between the three coordinates follow the order $\left|\boldsymbol{x_a} - \boldsymbol{x_b}\right| < \left|\boldsymbol{x_{a'}} - \boldsymbol{x_b}\right|$. If a and b co-occur in document A, a in document A likely refers to coordinates $\boldsymbol{x_a}$, rather than to $\boldsymbol{x_{a'}}$. Because the DF does not account for such ambiguities, the RS is expected to perform better than the DF in extracting spatially significant objects.

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

3) Regional Co-occurrence Variation (RV)

Underlying hypothesis: Landmarks are frequently mentioned in web documents together with a wide variety of place names in its surrounding.

The regional co-occurrence variation (RV) is another measure based on the co-occurrences between the target place name and its surrounding place names. Instead of using the total number of co-occurrences, the diversity in co-occurrences was used. The formula for the RV is as follows.

$$v(p_i) = |\{p_j \in P(p_i) | \kappa(p_i, p_j) \geq 1\}|$$

As with the RS, $P(p_i)$ is the physical proximity of the target place name p_i, and $\kappa(p_i, p_j)$ is the number of co-occurrences between place names p_i and p_j.

4) Spatial Sentence Frequency (SF)

Underlying Hypothesis: Landmarks are frequently mentioned in spatial sentences.

The spatial sentence frequency (SF) represents the frequency that the target place name is used in sentences that discuss spatial subjects. We estimate here that a sentence containing both a place name and also a *spatial trigger phrase* discusses spatial subject. We manually created a set of spatial trigger phrases. The formula for the SF is as follows.

$$s(p_i) = |\{d \in D | p_i \in s \wedge e \in s \wedge s \in d \wedge e \in E\}|$$

Here, D is the set of documents, d is a document, s is a sentence, p_i is the target place name, E is the set of spatial trigger phrases, and e is a spatial trigger phrase. Table 3 lists some of the spatial trigger phrases used in the extraction.

Table 3. Examples of spatial trigger phrases

Actions	walk, drive, turn at, go up, go down, arrive, stop at
Directions	right, left, front, back
Orientation	north, south, east, west
Spatial relationships	next to, at the corner of, behind, other side of
Spatial objects	intersection, road, street, railroad crossing
Means of transportation	car, bicycle, train

5) Case Frequency (CF)

Underlying Hypothesis: Landmarks are frequently used in spatial deep structure cases.

The case frequency (CF) focuses on the *case* that a place name accompanies. According to Fillmore's case grammar, each noun phrase in a sentence belongs a certain deep structure case, which means a specific role assigned to the noun phrase [11].

Some of the examples of deep structure cases are a *subject*, an *object*, a *location*, a *source*, a *method*, and a *goal*. In case grammar, the predicate is considered to be the central element in a sentence. The subject, the direct object, the indirect object, and prepositional phrases are all considered as noun phrases that modify the predicate. The deep structure cases are sometimes used for information retrieval [36].

Because a deep structure case indicates the phrase's role in a sentence, the frequency that the target place name is used in a spatial case is speculated to reflect the significance of the object in a spatial role.

Although the underlying deep structure of cases is common to all natural languages, the surface structure may vary. In isolated or inflective languages such as English and most other Indo-European languages, the case is expressed either by a preposition or word order. On the other hand, in agglutinative languages such as Japanese, Korean, Hungarian, and Finnish, the case is expressed by a suffix or a case particle.

One of the most common styles of spatial description in Japanese is as follows.

$$(w*) + pn + cp + (w*) + sp$$

Here, w is a term in general, pn is a place name, cp is a case particle, sp is a spatial predicate, and $*$ indicates an arbitrary number of repetition.

Because our target area is a city in Japan, we used case particles as the indicators of a deep structure case. We selected a set of Japanese case particles that often indicate spatial deep structure cases: *kara, made, yori, e, ni*, and *de*, which roughly correspond to the English prepositions *from, until, from, toward, to*, and *at*, respectively.

We define the case frequency (CF) as the frequency where the target place name is followed by the spatial case particle. The formula for the CF is as follows.

$$c(p_i) = |\{d \in D | p_i \in s \wedge c \in s \wedge s \in d \wedge c \in C_+ \wedge \alpha(p_i, c)\}|$$

Here, D is the set of documents, d is a document, s is a sentence, p_i is the target place name, C_+ is the set of case particles indicating a spatial deep structure case, c is a case particle, and $\alpha(p_i, c)$ indicates adjacency between p_i and c within a sentence (defined as true if p_i and c appear in this order).

5 Experiment

We performed a series of experiments to compare the validity of the proposed measures of landmark significance. In the experiment, we asked subjects to name a set of place names that they consider to be the landmarks of the target area. We compared the human judged sets of landmarks with the aforementioned five measures in terms of the recall and precision. This is a common evaluation method in information retrieval [29]. The recall and precision curves were graphed for the sets of landmarks extracted from the GIS data based on our five measures.

5.1 Data Set

The data set used in the experiment is as follows.

Subjects: 50 subjects consisted of 36 residents of the target area, Kyoto, and 14 people from outside the city. 40 were male and 10 were female.

Answer Set: Each subject was asked to name 20 of the most notable landmarks in Kyoto. A total of 1,000 entries consisted of 275 different place names. Table 4 lists the most frequently mentioned place names.

Table 4. Top 10 significant landmarks in Kyoto, collected from the subjects

Place name	# of answers
Kinkakuji (Golden Pavilion)	44
Ginkakuji (Silver Pavilion)	43
Kiyomizudera Temple	42
Kyoto Station	39
Kyoto Tower	34
Heian Shrine	32
Kyoto Imperial Palace	30
Kyoto University	29
Nijo Castle	29
Yasaka Shrine	25

GIS Data: The five measures for the cognitive significance were applied to the place names taken from a regular GIS, a digitized residential map provided by Zenrin, Ltd. [38]. This map data is divided into layers, including a "significant objects" layer that contains 7,109 place names. Although we can assume objects in this layer are mostly potential landmarks, their levels of significance vary. Famous temples and ordinary elementary schools alike are included in this layer. Thus, our goal in this experiment was to assign the level of significance to each of the place names included in the "significant objects" layer.

Web Documents: We collected 157,297 regional web pages for the web documents that were used to calculate our measures. Only the text part was used in the information extraction. The total file size was 2.45GB.

A focused web crawler was used for the collection of web pages. A focused web crawler is a special type of crawler that collects only the pages meeting a certain criterion [4,6]. The links are traced only when a page satisfies the criterion. In many cases, focused crawlers have greater efficiency in retrieving web pages under a certain topic, than regular web crawlers do. In this experiment, we used the place names taken from the GIS as the criterion of collection. Each page was guaranteed to contain at least one place name in the target area. The details of our implementation of a focused crawler are described in our previous paper [32].

Figure 7 is the architecture of the system that we implemented to evaluate our measures.

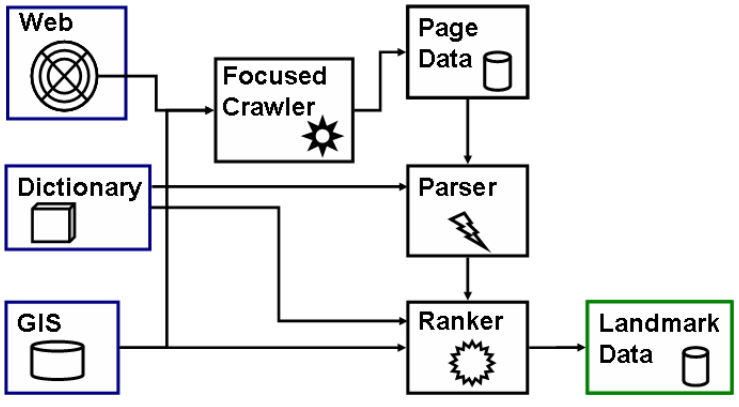

Fig. 7. System architecture for the measurement system

5.2 Evaluation Method

The definitions of the precision and recall are as follows.

$$\text{Precision} = \frac{\text{Retrieved Correct Objects}}{\text{Retrieved Objects}}$$

$$\text{Recall} = \frac{\text{Retrieved Correct Objects}}{\text{Correct Objects in the Population}}$$

When the rank position k is small, the set is likely to have high precision and low recall, while when k is large, the set tends to be the opposite. The precision and recall are functions of the rank position k. A precision-recall curve (P-R curve) is commonly used to visualize a series of the precision and recall pairs obtained by altering k [26].

The place names from the GIS in our evaluation were set in a decreasing order of the values calculated by each measure (DF, RS, RV, SF, and CF). The top k place names were selected, and the precision and recall were calculated with respect to a set of human-judged landmarks.

We consider only the points where the value of the recall increases to make the P-R curve smooth. By re-numbering the extracted pairs, we obtained the series of P-R pairs as a function of a new parameter j.

Then, we averaged the P-R pairs collected from different subjects for each j and obtained the averaged P-R curve, which is a function of j. This is called averaging by micro-evaluation [26]. In our case, the k ranged from 1 to 7,109 (= the number of the potential landmarks in GIS), j ranged from 1 to 20 (= the number of the "correct landmarks" given by each subject), and the number of the P-R pair series (to be averaged) was 50 (= the number of the subjects).

The evaluator consisted of a PostgreSQL database and approximately 1,100 lines of Perl scripts, including the part where each measure is calculated.

5.3 Results

Figures 8-11 indicate the comparison of the averaged P-R curves for the five measures. In these figures, RS, RV, SF, and CF, which uses the spatial context, are each compared with the DF, which does not employ the spatial context.

Table 5 compares precisions of five measures for different rank positions.

5.4 Discussion

The results of the experiments showed that in the overall performance the measures with spatial context (RS, RV, SF, CF) matched better with the human-judged sets of landmarks, in comparison to a measure without spatial context (DF).

The regional co-occurrence summation (RS) gave especially high precision for low recall situations, which means that the RS is the best measure to use when only the

Table 5. Precision of each measure

P. at rank	DF	RS	RV	SF	CF
5	0.016	0.368	0.132	0.272	0.152
10	0.140	0.212	0.186	0.276	0.130
15	0.096	0.259	0.245	0.241	0.147
20	0.106	0.239	0.192	0.251	0.152

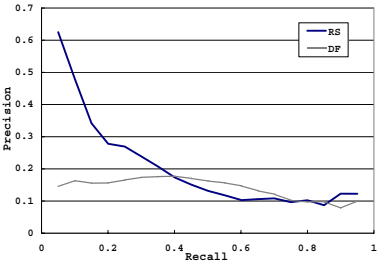

Fig. 8. RS and DF

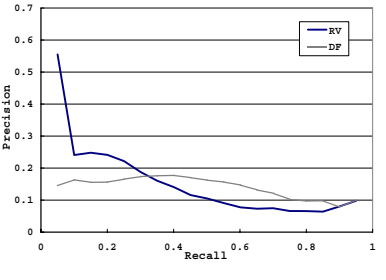

Fig. 9. RV and DF

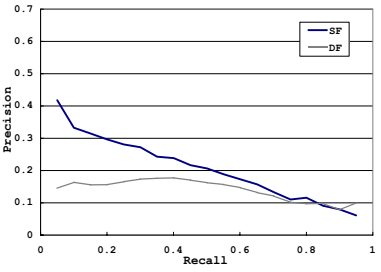

Fig. 10. SF and DF

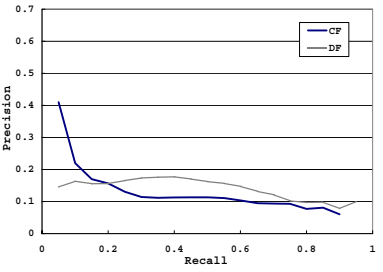

Fig. 11. CF and DF

most significant landmarks are required. These are the cases showing only the typical landmarks of the region, such as in case of roughly abstracted route maps.

On the other hand, the spatial sentence frequency (SF) had relatively high precision for high recall situations. If a large set of landmarks is required, the SF is the preferred measure. For sightseeing maps where users want to see a large number of landmarks, the SF can be used.

In calculating the RS, all neighboring place names were treated equally. However, we could have used heterogeneous data sources, for example in different scales, and calculated the RS in a multi-layered manner. This approach may help avoid ambiguities in place names and improve the result.

For low recall situations, the linguistic approaches (SF and CF) did not perform as well as the statistical approaches (RS and RV). This is probably because of the relative scarcity of the obtained samples in comparison with the statistical approach.

Since linguistic approaches aggregates longer patterns of terms than the statistical approaches, with smaller size of documents, the signal-to-noise ratio increases. An increase in the number of collected web pages may change the situation better for the linguistic approaches.

Although using a specific measure proposed in this paper is the simplest solution for presenting landmarks, a combined measure could be used also.

6 Conclusion

In this paper, we discussed extracting cognitively significant geographic objects using web documents. Our main achievement is that we generated new geographic knowledge that is not present in conventional GIS, by aggregating distributed region related document on the Web.

Our approach has also introduced spatial context into text mining. The results of experiments showed that measures adopting spatial context match with human judged landmarks better when compared with the document frequency (DF).

In this paper, we did not consider how the effect of spatial tasks held by different users to the significances of landmarks. Such personalization is a difficult task, yet we are considering the extraction of such information by focusing on the grammatical aspects of the spatial sentences. For example, the types of the cases may provide clues. In calculating the CF, we only used the total frequency of spatial cases. However, their

types are closely related to the spatial tasks. An analysis on the spectrum of the cases may unravel the preponderant spatial task for each landmark.

One advantage of our proposed method is that it can measure the significance of landmarks *quantitatively*. Although human map editors can choose significant landmarks when creating maps, giving each object its *level of significance* is often a difficult question.

The proposed method can be used in applications such as a progressively zoomable map interface, since place names shown on the map must be altered as the scale changes. The system can show the user the most significant and important place names on the map interface, without cramming too many characters on the screen.

Unlike questionnaire-based methods that are too expensive for collecting answers and analyzing results, our method is *scalable*, i.e. it can be extended simply by collecting more web pages. Because the size of the Web is growing continuously, the precision of our method is speculated to rise.

References

1. G. E. Burnett, D. Smith, and A. J. May, Supporting the navigation task: characteristics of 'good' landmarks, *Proceedings of the Annual Conference of the Ergonomics Society*, Taylor & Francis, 2001
2. O. Buyukokkten, J. Cho, H. Garcia-Molina, L. Gravano, N. Shivakumar, Exploiting geographical location information of web pages, *Proceedings of Workshop on Web Databases (WebDB99) held in conjunction with ACM SIGMOD99*, pp. 91-96, 1999
3. C. Brenner and B. Elias, Extracting landmarks for car navigation systems using existing GIS databases and laser scanning, *Proceedings of the ISPRS Workshop 'Photogrammetric Image Analysis,'* Munchen, Germany, 2003
4. S. Chakrabarti, M. van den Berg, B. Domc, Focused crawling: a new approach to topic-specific web resource discovery, *Proceedings of the 8th International World Wide Web Conference (WWW8)*, Toronto, Canada, 1999
5. H. Couclelis, R. Golledge, N. Gale and W. Tobler, Exploring the anchor-point hypothesis of spatial cognition, *Journal of Environmental Psychology*, Vol. 7, No. 2, pp. 99-122, 1987
6. M. Diligenti, F. M. Coetzee, S. Lawrence, C. L. Giles and M. Gori, Focused crawling using context graphs, *Proceedings of the 26th International Conference on Very Large Databases (VLDB 2000)*, Cairo, Egypt, 2000
7. R. M. Downs and D. Stea (Eds.), *Image and Environment : Cognitive Mapping and Spatial Behavior*, Aldine Publishing Co., Chicago, Illinois, 1973
8. M. J. Egenhofer and D. M. Mark, Naive geography, A. U. Frank and W. Kuhn (Eds.), *Spatial Information Theory: A Theoretical Basis for GIS*, Lecture Notes in Computer Science 988, pp. 1-15, Springer-Verlag, Berlin, 1995
9. B. Elias, Extracting landmarks with data mining methods, W. Kuhn, M. Worboys, and S. Timpf (Eds.), *Spatial Information Theory: Foundations of Geographic Information Science*, Lecture Notes in Computer Science 2825, Springer-Verlag, pp. 375-389, 2003
10. B. Elias and C. Brenner, Automatic Generation and Application of Landmarks in Navigation Data Sets, P. Fisher, (Eds.), *Developments in Spatial Data Handling*, pp. 469-480, Springer-Verlag, Berlin, 2004
11. C. J. Fillmore, The case for case, E. Bach and R. T. Harms (Eds.), *Universals in Linguistic Theory*, Holt, Rinehart and Winston, Inc., pp. 1-88, 1968
12. P. Gould and R. White, *Mental Maps*, Pelican Books, 1974

13. K. Koiso, T. Mori, H. Kawagishi, K. Tanaka, and T. Matsumoto, InfoLOD and LandMark: spatial presentation of attribute information and computing representative objects for spatial data, *International Journal of Cooperative Information Systems*, Vol. 9, No. 1-2, pp. 53-76, 2000
14. F. Lee, S. Bressan, and B. Ooi, Hybrid transformation for indexing and searching Web documents in the cartographic paradigm, *Information Systems*, Vol. 26, No. 2, pp. 75-92, 2001
15. K. Lynch, *The Image of the City*, MIT Press, Cambridge, Massachusetts, 1960
16. D. M. Mark and A. U. Frank, Experiential and formal models of geographic space, *Environment and Planning*, Vol. 23, No. 1, pp. 3-24, 1996
17. K. S. McCurley, Geospatial mapping and navigation of the web, *Proceedings of the Tenth International World Wide Web Conference (WWW10)*, pp. 221-229, Hong Kong, 2001
18. Metacarta : geographic text search engine, http://www.metacarta.com/technology/index.html#geoparse
19. P. E. Michon and M. Denis, When and why are visual landmarks used in giving directions?, D. R. Montello (Ed.), *Spatial Information Theory: Foundations of Geographic Information Science*, Lecture Notes in Computer Science 2205, Springer-Verlag, pp. 292-305, 2001
20. G. A. Mitra, C. Buckley, A. Singhal and C. Cardie, An analysis of statistical and syntactic phrases, *Proceedings of 5th International Conference on Computer-Assisted Information Searching on Internet (RIAO'97)*, pp. 200-214, Montreal, Canada, 1997
21. I. Moon, J. Miura and Y. Shirai, Automatic extraction of visual landmarks for a mobile robot under uncertainty of vision and motion, *IEEE International Conference on Robotics and Automation*, pp. 1188-1193, 2001
22. U. Neisser, *Cognition and reality: principles and implications of cognitive psychology*, W. H. Freeman and Company, San Francisco, 1976
23. C. Nothegger, S. Winter and M. Raubal, Selection of Salient Features for Route Directions, *Spatial Cognition and Computation*, Vol. 4, No. 2, pp. 113-136, 2004
24. *The Oxford English Dictionary*, Second Edition, Oxford University Press, 1989
25. M. Raubal and S. Winter, Enriching wayfinding instructions with local landmarks, M. Egenhofer and D. Mark (Eds.), *Geographic Information Science*, Lecture Notes in Computer Science 2478, Springer-Verlag, pp. 243-259, 2003
26. C. J. van Rijsbergen, *Information Retrieval - Second Edition*, Butterworth & Co. Publishers Ltd, 1979
27. D. E. Rumelhart, *Introduction to Human Information Processing*, John Wiley & Sons, Inc., 1977
28. T. Sagara, M. Arikawa, and M. Sakauchi, Spatial Information Extraction System Using Geo-Reference Information, *Information Processing Society of Japan Journal:Database, Vol. 41, No. SIG6 (TOD7)*, pp. 69-80, 2000
29. G. Salton, *Automatic Information Organization and Retrieval*, McGraw-Hill Inc., 1968
30. M. Sorrows and S. Hirtle, The nature of landmarks for real and electronic spaces, C. Freska and D. Mark (Eds.), *Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science*, Lecture Notes in Computer Science 1661, Springer-Verlag, pp. 37-55, 1999
31. T. Strzalkowski, Natural language information retrieval, *Information Processing and Management*, Vol. 31, No. 3, pp. 397-417, 1995
32. T. Tezuka, R. Lee, H. Takakura, and Y. Kambayashi, Integrated model for a region-specific search system and its implementation, *Proceedings of the International Conference on Internet Information Retrieval (ICIIR 2003)*, pp. 243-248, Koyang, Korea, 2003
33. T. Tezuka, Y. Yokota, M. Iwaihara, and K. Tanaka, Extraction of cognitively-significant place names and regions from web-based physical proximity co-occurrences, X. Zhou, S. Su, M. P. Papazoglou, M. E. Orlowska, and K. G. Jeffery (Eds.), *Web Information Systems - WISE 2004*, Lecture Notes in Computer Science 3306, pp. 113-124, Springer-Verlag, 2004

34. A. Tom and M. Denis, Referring to landmark or street information in route directions: what difference does it make?, W. Kuhn, M. Worboys, and S. Timpf (Eds.), *Spatial Information Theory: Foundations of Geographic Information Science*, Lecture Notes in Computer Science 2825, Springer-Verlag, pp. 375-389, 2003
35. C. S. Yadav (Eds.), *Perceptual and Cognitive Image of the City*, Concept Publishing Company, New Delhi, India, 1987
36. E. B. Wendlandt and J. R. Driscoll, Incorporating a semantic analysis into a document retrieval strategy, *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 270-279, Chicago, 1991
37. A. G. Woodruff and C. Plaunt, GIPSY: automated geographic indexing of text documents, *Journal of the American Society for Information Science*, Vol. 45, No. 9, pp. 645-655, 1994
38. Zenrin Co.,Ltd, http://www.zenrin.co.jp/