Preliminary Workshop on Evaluation of Geographic Information Retrieval Systems for Web Documents

Masatoshi Arikawa Takeshi Sagara Kouzou Noaki Hideyuki Fujita
Center for Spatial Information Science, The University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo 153-8904 Japan
arikawa@csis.u-tokyo.ac.jp

Abstract

Geographic Information Task (GeoTask) is one of the newly proposed tasks at the NTCIR-4 WEB. Geographic information is close to our daily lives, and is one of the real ways to access Web information. Researches and developments of such aspects have been increasing recently, however, comparative evaluations of such kinds of techniques has not been carried out so far. GeoTask focused on the technology that the system extracts geographic information from Web documents relevant to a given viewpoint. The aim of this workshop is to expedite and advance researches and developments of Geographic IR technologies for the Web, therefore we are going to build reusable test collection for evaluating various methods of Geographic IR for Web documents. In this paper, challenges for searching geographic information are described, which are discussed through the explanation of our research on developing geographic IR systems.

Keywords: *Web, Information Retrieval, Geographic Information.*

1. Introduction

It is useful to access Web pages through geographic keys which are connected to locations in the real world, such as positions, addresses, telephone numbers and so on. We have organized "Geographic Information Task1 (GeoTask)" as a subtask of NTCIR4-WEB, and made a test collection of GeoTask. However, we have not received any result submissions for the subtask. We have had great experience from preparing the test collection and have thought the reasons of why there is no result submission. This report discusses the reasons. We hope that the discussion will help make the next test collections for geographic information retrieval.

This paper first introduces a test collection of GeoTask. Then, we discuss our current research on geographic information retrieval for Web documents which are not part of NTCIR4-WEB document collection. We intend to find real useful applications using NTT Yellow Pages and map data. Our research would be useful to make the next test collections from the practical viewpoint.

2. Overview of GeoTask

We, the organizer of GeoTask, made a test collection for geographic information retrieval based on NTCIR4-WEB document collection.

2.1. Making target dataset

NTCIR4-WEB document collection was considered too large for GeoTask participants to concentrate on essential research works on GeoTask. Thus, we made a small set of dataset that is called "the target dataset" of GeoTask. It was supposed to relieve the participants of the burden of developing the techniques of treating a large amount of Web documents. Each document of the target dataset selected from NTCIR4-WEB document collection must satisfy the following two conditions:

- The document must include a geographic name in Tokyo area. It means that geographic information in Tokyo was supposed to be used in our prepared questions as the test collection of GeoTask.
- The document must include one of the keywords which were supposed to be used in our prepared questions.

The above selection was done using a morphological analysis tool "Chasen" and geographic name dictionaries covering Tokyo area. The number of documents of our target dataset is approximately 240,000.

2.2. Questions as parts of the test collection

Figure 1 shows an example of questions in the test collection of GeoTask. This example means the query to extract the information of all universities in Tokyo area. The area of the query is "Tokyo". Its keyword is "university". The result of the question is supposed to a set of a tuple of (1) document id, (2) name of university, address of university and geographic coordinates of university. The organizer prepared several questions for the test collections (Figure 2).

```
[QUESTION]
<GEO>
  <NUM>0001</NUM>
  <SPAQ TYPE = "AREA" OUTPUT = "POINTS">
    <KEYWORD>大学</KEYWORD>
    <LOCALITY TYPE="REF">東京都</LOCALITY>
  </SPAQ>
  <DESC>東京都内の大学に関する地理情報を抽出せよ.
  </DESC>
</GEO>

[RESULT]
NW000003291, 東京大学, 文京区本郷 7-3-1,
    139.768616,35.708927
NW000004193,東京都立医療技術短期大学, 03-3819-1211,
    139.77603, 35.74796
NW000091353,東京医科歯科大学,営団丸ノ内線お茶の水駅
    徒歩1分, 139.76750, 35.69838
.....
```

Figure 1. Examples of question and result.

2.3. Data and query processing for questions

There are basic three processing phases of Web documents in GeoTask.

(1) Analyzing Web documents
(2) Geographic information extraction from Web documents
(3) Geocoding: converting indirectly location reference descriptions into directly location reference descriptions.

Also, there is query processing which depends on approaches of solving given queries.

(4) Query processing

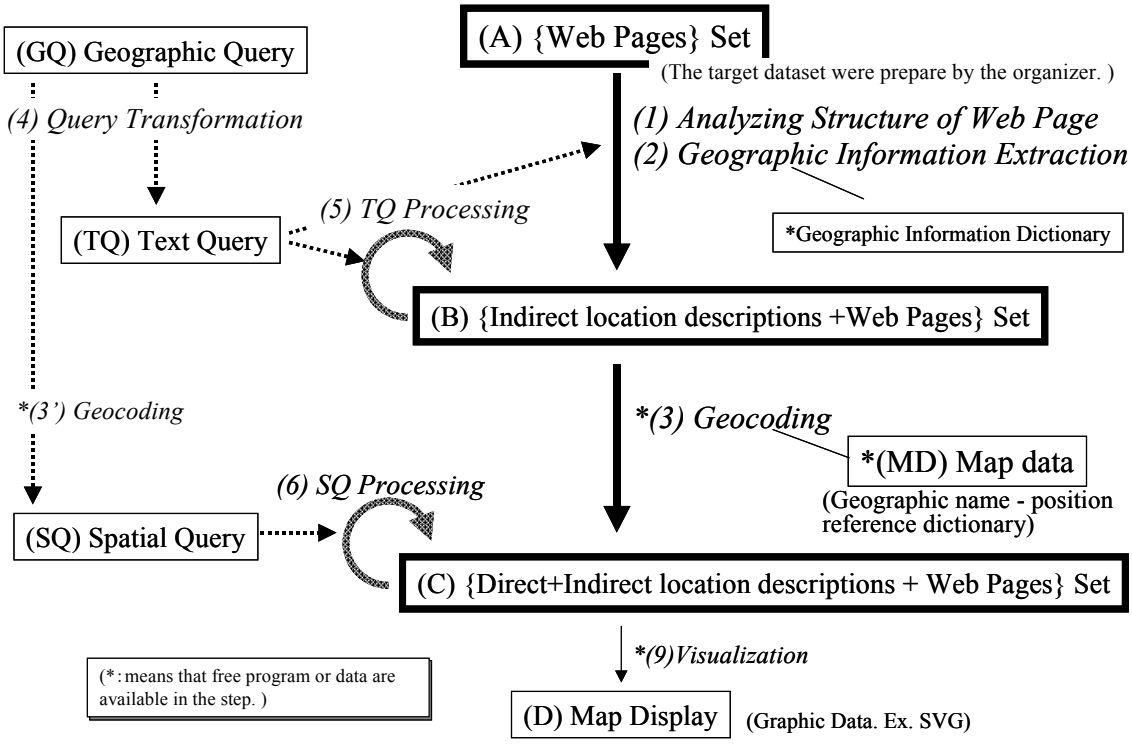

Figure 2. Basic processing flow of geographic information retrieval.

2.5. Analysis of Web documents

One document may be composed of multiple themes. For example, there is a Web page including a list of links. One Web page can have more than ten restaurants as components. In this case, one Web page should be considered more than ten logical Web pages, because one logical page should have only one theme. HTML documents do not have explicit structures for representing multiple themes in the contents. It is necessary to develop the methods of understanding and analyzing the structures using experimental tools. There have been known research works related to analyzing structures from the viewpoint of geographic information. The research work focuses on anchor tags, heading tags and alt attributes in image tags for finding structures.

2.6. Geographic information extraction

There are two kinds of geographic descriptions. First one is “directly location reference descriptions”. Examples of directly location reference descriptions are longitude and latitude, and the coordinates for a survey and a map. The other one is “indirectly location reference descriptions”. Examples of the indirectly location reference descriptions are address, land names, landmark names, postal code, telephone number, IDs for pole on streets, and expressions of route guides. We also should make the difference between human readable and machine readable for indirectly location reference descriptions. IDs and numbers such as postal code and telephone numbers were designed for machine readability. On the other hand, addresses and land names are designed for human readability. Both machine and human readable descriptions have good and bad points, or their characteristics. Good and bad points of human readable indirectly location reference descriptions may have some variations and abbreviations. The characteristics are good for human, but bad for computers. Geographic information extraction is the processing of finding appropriate geographic descriptions from Web documents. There may be multiple candidates of geographic descriptions in one Web page. It is essential to decide which geographic descriptions are relevant to the Web page.

2.7. Geocoding

Geocoding is the processing of converting indirectly location reference descriptions into directly location reference descriptions. Geocoding must use geographic dictionaries or map data which have information of corresponding relations between indirectly location reference descriptions and directly location reference descriptions. Main problems of geocoding are efficiency and robustness. Human readable geographic information may have ambiguity. Geocoding should have tolerance for the ambiguity.

We, organizer have studied on geocoding for a few years and developed efficient and robust software tools for geocoding. We planed to provide participants with the geocoding tool as ASP (application service provider) of the service of the Center for Spatial Information Science at the University of Tokyo. We considered that the part of geocoding process is out of the range of the GeoTask test collection. The competition on GeoTask test collection should be done in the processing of geographic information extraction.

2.8. Query transformation

Questions include conditions of both spatial and textual queries. There may be at least two typical approaches:

(1) Textual query processing unification
All spatial queries are converted into textual queries. For example, the coordinates are converted into geographic areas or positions. All spatial queries are converted into and unified with textual queries.

(2) Spatial query processing unification
This is the opposite way of (1). All textual queries concerning locations and areas can be converted into spatial queries. For example, the word "Tokyo station" could be converted into the coordinate (x,y) as the central point of the Tokyo station. Thus, all Web documents could be bounds with locations, areas or lines as additional spatial information about the Web documents for spatial queries.

2.9. Evaluation method

We planed to adopt the pooling method for evaluate the result submissions for GeoTask test collection. We proposed three evaluation criteria. The first and second ones are precision and recall rates. The third one is original one, that is, "distance" between the location as an answer and the right location as a correct answer. The result submissions include the coordinates for geographic information extracted from Web documents. The coordinates can be visualized as a map or maps easily. In other words, the result submission of GeoTask can be considered as maps. The third evaluation criterion means distance errors between result maps and right answer maps.

2.10. Research tasks

GeoTask is the first trial. It was difficult to decide an appropriate test collection. Points of the difficulty are as follows.

- removing unnecessary documents
- precision or recall intense selection
- making right answer set or interpretation of right answer set

2.11. Provided data and tools

GeoTask test collection provided the information about free geographic information dictionaries, free map data, our developed geocoding and visualization tool.

3. Collecting and scoring spatial documents

In this section, we discuss an efficient algorithm for retrieving spatial documents from the Web. Spatial documents are documents having location information as part of the content. Then, we propose scoring methods for sorting the documents. Finally, a prototype system of spatial search engine based on our proposed methods is shown.

3.1. Crawling spatial documents

The basic algorithm for adding spatial index to spatial documents is quite simple; (a) Parse the document and extract location names. (b) Convert those names into coordinate values such as longitude and latitude. The process (a) is called "geoparse" and (b) is called "geocode"[2]. These two processes are the most important technologies to exploit spatial documents.

We had already developed practical geoparser and geocoder for Japanese spatial documents. The system was reported as a "spatial document management system (SDMS)" in 2001[3]. The system can add street-block-level spatial index in almost all part of Japan to the documents written in Japanese.

Using SDMS, user can retrieve web documents by combination of spatial query such as range retrieval and full text search. For example, you can ask the system like "Show me all documents which are geo-referenced within 1km distance from Shinjuku station and contain 'restaurant'". However, there are two big issues remained to answer this question.

The first issue is how to retrieve web pages which will be geo-referenced in a certain area. Since it is not practical to collect and add spatial index to all web pages in the world, an efficient strategy is required to find spatial documents from the web.

The second issue is how to select documents which satisfy both spatial and keyword conditions. Even though a document is geo-referenced in the queried area and contains queried keyword such as 'restaurant', it does not ensure that the document describes "restaurant in the area", because the document may mention about an office in the area and a restaurant outside of the area. Confirming relevancy between the location name and the keyword essentially requires semantic text understanding.

Furthermore, we should consider that the keyword often represents a concept or a category in geographic information retrieval. In the restaurant query, the user would like to obtain the list of restaurants and their documents, but not the list of documents including a word "restaurant". In our experience, some web documents introduce restaurants as bar or pub, and they do not contain a word "restaurant" in their text, thus these documents cannot be retrieved by the query.

Therefore, we utilize the Yellow Pages for collecting web pages. The Yellow Pages contain almost all shops with their name, postal address, phone number and category of business. The crawling algorithm is shown below.

Algorithm 1: Crawling spatial documents.

Step 1. Pick up a record $\boldsymbol{y}$ from the Yellow Pages.
Step 2. Make search keyword from name, address, phone number in $\boldsymbol{y}$.
Step 3. Calculate coordinate values $\boldsymbol{g}$ of $\boldsymbol{y}$ by geocoding the address.
Step 4. Collect web documents $\boldsymbol{d}_{1..m}$ ($\boldsymbol{m}$: maximum pages) using keyword index.
Step 5. Check $\boldsymbol{d}_i$, $\boldsymbol{i}$ = 1 .. $\boldsymbol{m}$ and store ($\boldsymbol{y}$, $\boldsymbol{g}$, $\boldsymbol{d}_i$) to the relational database if
a. $\boldsymbol{d}_i$ contains correct name and address, or
b. $\boldsymbol{d}_i$ contains correct phone number
Step 6. Go back to Step 1 while more record remains.

The reason why the document contains correct name will not be stored in step 5 is, it may describes another shop in another place with the same name.

This algorithm is relatively fast hence it uses normal keyword index search. We implemented and applied the algorithm actually for about 100 thousand restaurants in Tokyo area and collected about 450 thousand spatial documents (which satisfy one of the conditions in step 5.) in 4 days. The documents can be considered that they really describe $\boldsymbol{y}$ and their locations can be referenced by $\boldsymbol{g}$, therefore the relevancy between the location name and the keyword does not become a problem. Additionally, the documents can be retrieved by categories of business stored in $\boldsymbol{y}$.

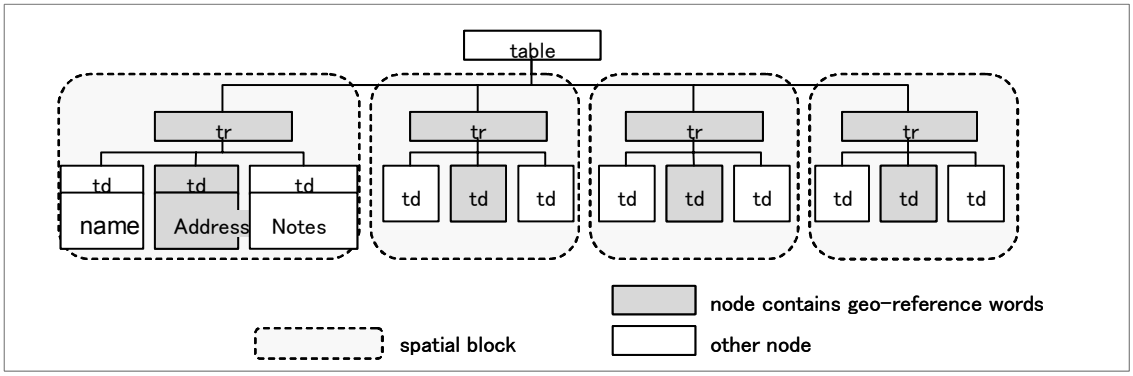

Figure 3. Extracting spatial blocks from spatial document.

3.2. Scoring spatial documents

Scoring documents is one of the most important technologies of information retrieval (IR), since user would not like to read all documents. In geographic information retrieval, we divided scoring method into two different levels. The first scoring method evaluates "popularity" of each geographic object, and second one evaluates "quality" of each document.

For example, when a user asks about "restaurants near to the station", the answer should be a list of restaurants sorted by some measures of qualities, such as taste of the meals served at the restaurant or reputation of the restaurant. Hence such subjective information is too difficult to extract from the documents, popularity is used alternatively. Conceptually, if there is more number of web documents describing restaurant A than restaurant B, the restaurant A can be thought as more popular on the web than B. More precise definition will be shown at the end of this section.

When the user gets interest in restaurant A, he/she would like to examine by reading spatial documents related to the restaurant. In this case, reliable and informative documents should be presented with high priority. To determine the reliability of web pages, some techniques, which had been already developed based on link relationships such as page rank, can be applied. In our implementation, reliability can take binary value, i.e. 'reliable' and 'not reliable'. When the page rank of a document is higher than the threshold, it is considered as reliable. Additionally, we confirmed by an evaluation experiment that documents containing both correct phone number and address are reliable with high possibility. Therefore, when the document contains phone number and address, it considered reliable regardless of its page rank.

To evaluate whether the document is informative or not, the number of characters is simply used. However, since some web documents are describing multiple geographic objects sequentially in one page, such as a list of the author's favorite restaurants, sentences which really mention about the object must be extracted before counting characters. Avoiding text understanding, we use a heuristic approach to implement the extraction process. First, "spatial block" is defined as a smallest part of web document which contains more than one spatial content, and the block is segmented by any HTML tag. The extraction is done by the block. Second, two assumptions are made.

1. A spatial block contains more than one location name.
2. Spatial blocks in a web page are placed in parallel under certain tag-block.

The spatial block can be extracted by the algorithm shown below (Figure 3).

Algorithm 2: Spatial block extraction.

Step 1. Create HTML tag tree from the web document.
Step 2. Mark every node which text contains location names.
Step 3. Mark every parent node of marked node repeatedly.
Step 4. Find most upper level nodes which are marked, and its brother node is also marked. Every marked node in the level is spatial block.
Step 5. Extract the spatial block which contains location name focused on.

Although the algorithm can work only when the blocks have location names and segmented by HTML tags, it is so flexible that it can extract either a line surrounded by '<tr>' and '</tr>' tags from a

table, or a block surrounded by ' ' and ' ' (Fig. 1).

The quality of a document ***q*** is defined by a combination of reliability ***r*** and the number of characters in the spatial block ***n*** as;

1. if ***r*** = 'reliable' and ***n*** >= ***th*** then ***q*** := 1.0
2. if ***r*** = 'reliable' and ***n*** < th then ***q*** := ***n*** / ***th***
3. if ***r*** = 'not reliable' and ***n*** >= ***th*** then ***q*** := ***w***
4. if ***r*** = 'not reliable' and ***n*** < ***th*** then ***q*** := ***w*** (***n*** / ***th***), where ***th*** is a threshold value, ***w*** is a weight ($0 < w < 1.0$).

And the popularity ***p*** of a geographic object is defined as a total of ***q*** which is related to the object.

3.3. A prototype system of spatial search engine: restaurant search

Using proposed methods, a prototype system of spatial search engine was developed (Figure 4). The system can retrieve restaurants by a combination of its location, category of business and keywords. The search results will be shown by both a street map and a list of restaurants sorted by their popularity. By clicking a rectangle on the map or name of a restaurant, list of all web documents related to the restaurant will be shown sorted by their quality. The list contains links to their original pages, and texts from the spatial blocks.

4. Conclusions

We launched GeoTask, but we were unable to receive any result submissions for our proposed test collection. Some various levels of problems existed in setting this test collection as well as in the environment we provided for participants. The target questions should have been smaller than the ones we proposed. However, we did not have enough time to consider whether or not our proposed test collection is truly appropriate and feasible for expected participants. From this failure, we have gained many experiences, and now we discuss better proposals for the next GeoTask in the following paragraphs.

There are two major steps in geographic information retrieval. The first step is to extract geographic descriptions from Web documents. The next step is to extract meaningful or useful information about keywords other than the location information. This test collection of GeoTask covers both steps. It is wider for participants to solve the both steps. We should have divided our proposed tasks into more than two steps. Questions and results for the GoeTask test collection should be practically interesting and meaningful, but not for only the purpose of this test collection. To achieve the purpose, it is important to extract significant and useful information other than location information from Web pages. Also, it is a big issue whether these two keys, that is, keywords and locations, can be dealt with separately or not for GeoTask test collection from the practical viewpoint.

Then, we propose two concrete tasks for the next GeoTask as follows.

(1) Selecting spatial contents from Web documents

Spatial contents mean the contents describing some real-world objects. There is a Web document having geographic information, but is not a spatial content. For example, there is a sentence "I want to hold a live concert in a big place like in Hokkaido". This sentence has the location information "Hokkaido". However, this sentence does not directly describe about Hokkaido. We should remove such Web pages that are not directly related to the location in the real world from a set of geo-referenced Web documents. This problem can be solved by a classical natural language processing technique.

(2) Extracting significant and useful information from geo-referenced Web documents

In this task, participants are given a fixed theme and a small set of geo-referenced Web documents as spatial contents. Participants compete for finding or inducing more significant and useful information of high quality from the given geo-referenced Web documents. Examples of significant and useful information are opening hour and average price for most popular dishes in a restaurant. The task is theme oriented and may be solved using heuristic approach and machine learning.

References

[1] Geographic Information Task1 (GeoTask), NTCIR4-WEB, <http://smfp.csis.u-tokyo.ac.jp/~ntcir/> (In Japanese).

[2] Kevin S. McCurley, Geospatial Mapping and Navigation of the Web, ACM WWW10, Hong Kong, 2001.

[3] Sagara, T., Arikawa, M., Sakauchi, M., Spatial Document Management System Using Spatial Data Fusion, IIWAS2001, Linz, 2001, 399-409.

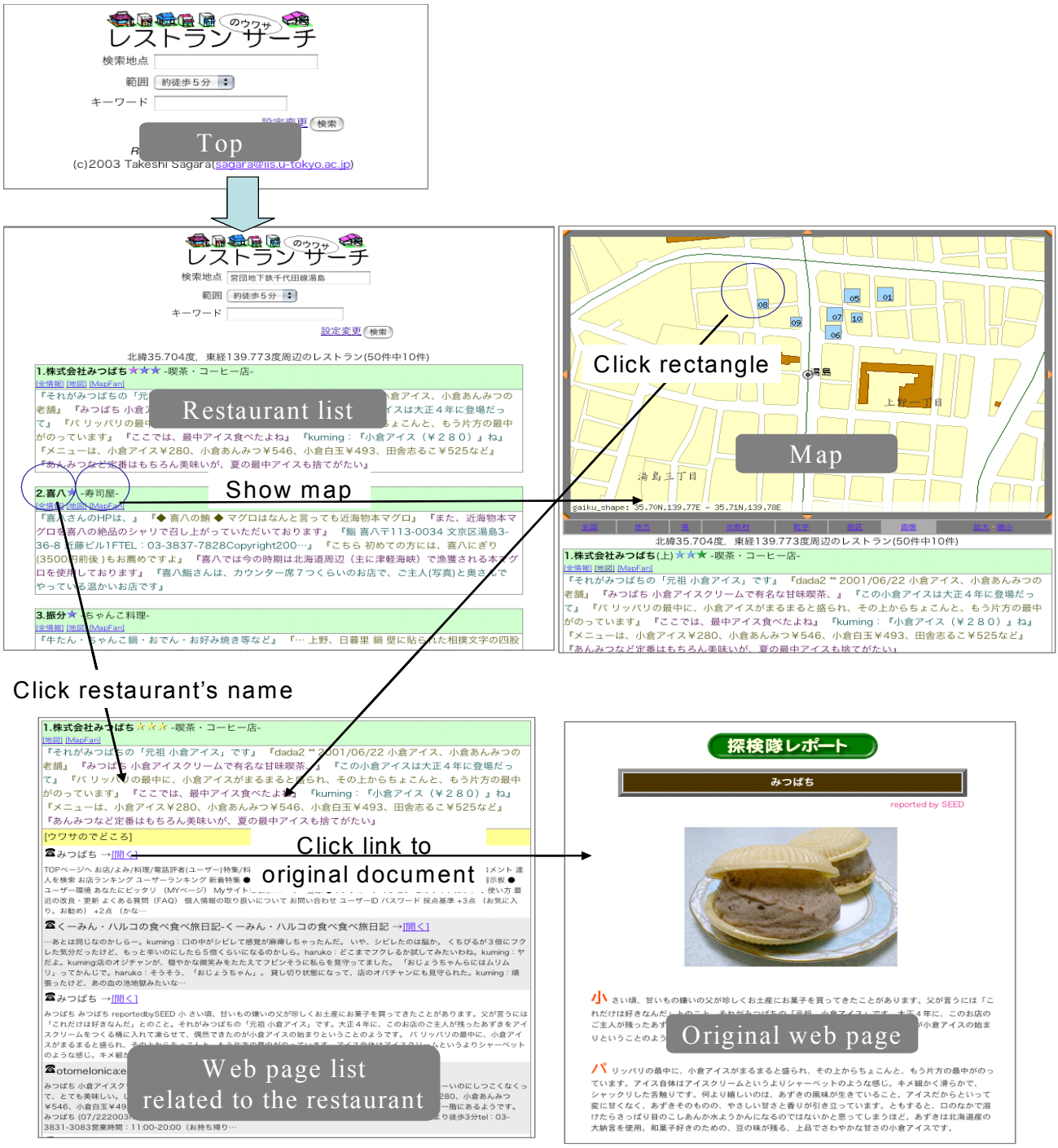

Figure 4. A prototype system of spatial search engine.