Fundamental Research of System for Extracting Attributes of Digital Map by Automatic Search on WWW

Monobe Kantaro[1], Tanaka Shigenori[2], Furuta Hitoshi [2], Kato Yuichi[3], and Nonaka Hiroshige[3]

Abstract: A service for providing spatial information is recently increasing. This service is easily used by a personal computer or a mobile phone. It is important that a digital map has a lot of attributes on the spatial service. However, making attributes of digital maps requires a great deal of time and money. Therefore, it is difficult to define attributes of spatial data and to produce these values. The purpose of this research is to develop a system that can automatically produce and update attributes of the digital map by searching on Web pages. Moreover, a function for searching spatial information with a keyword of natural language is implemented.

Keywords : GIS, Web, Digital Map, Attribute, Natural-Language Processing

1. Introduction

Recently, due to the development of information technology, importance of spatial information has been increasing significantly. Everyone is now able to acquire positional information easily, and spatial information has become indispensable in our daily life. Recently, services for spatial information have been increasing. People are able to use these services using a personal computer or a mobile phone. In order to further develop these services, maintenance of spatial information is essential.

Spatial information is composed of geometric information and attribute information. By the development of information processing technology, 3D expression of geometric information has been made possible. Most real-life-like digital maps have been constructed by the use of 3D technology. However, maintenance of attribute information is still undeveloped. For example, on digital maps that are available on the Web and on publication of the Geographical Survey Institute, there are a lot of cases where only minimum attribute information - such as building name, address, and telephone number - is maintained.

With this current situation, sufficient information will not be able to be offered to the expanding demands for services of spatial information in the future. For example, in the case of searching for "arch bridges that spans across Yodo river", if detailed attribute information such as "Yodo river" and "arch bridge" is not maintained in the digital map, the search will not be possible. Also, in the case of using GIS for disaster prevention, city planning, and barrier free transportation, detailed attribute information is necessary.

New information is in constant demand for attribute information. For providing accurate spatial information services, spatial information continuously requires[1)] to be in a state that corresponds to the real world. For this reason, attribute information needs[2)] to be frequently updated and maintained. Currently, maintaining attribute information requires great deal of cost and labor. Therefore, development of a system for easily maintaining attribute information is in demand. Hence, this research attempts to realize automatic extraction of attribute information on digital maps.

1) Student Member of JSCE, Master of Informatics, Faculty of Informatics Graduate Course, Kansai University
(2-1-1, Takatsuki-shi Ryouzenji-chou, Osaka 569-1095, JAPAN, E-mail: k_monobe@kb.kutc.kansai-u.ac.jp)
2) Member of JSCE, Dr. Eng., Professor, Faculty of Informatics, Kansai University
3) Non Member of JSCE, Bachelor of Informatics, Faculty of Informatics Graduate Course, Kansai University

2. Purpose of Research

For this research, we will primarily focus on information on the Web as data source of digital map's attribute information. Currently in Japan, 85.9 million Web pages exists[3]. By applying positional information contained in these Web pages and land objects on digital maps, we believe that extraction of information on the Web as attribute information will be possible.

In previous research by Sagara and his associates[4], research for effectively applying spatial information was performed by extracting original reference information that exists on the Web, and transforming the information to coordinate position. For the actual system, spatial information extraction system and spatial information search system was developed. In addition, embedding of spatial tag that uses XML expression has been proposed and is being evaluated. Nakajima and his associates[5] have constructed a map search system by using Web services. Service using SOAP messaging is provided, and map distribution is made possible. Saito and his associates[6] have researched geographic information search system that uses Semantic Web. Kubo and his associates[7] have developed GIS model system that unified spatial information and time-series information.

However, these researches have not come across automatic extraction of attribute information. Also, there are restrictions that place a limit to the Website. In addition, there is an information reliability problem due to extraction of imperfect positional information.

Therefore, this research attempts to develop a system that supports production of attribute information of digital map by automatically searching the WWW. We also plan to establish functions that search spatial information from natural language by using the extracted attribute information. Conceptual plan of this system is as shown in **Fig.1**. This system starts from extraction of attribute information from Web pages. Next, address matching is performed by using address information in the extracted attributes, and they are attached to land objects on digital maps. Also, a search of spatial information using natural language will be realized.

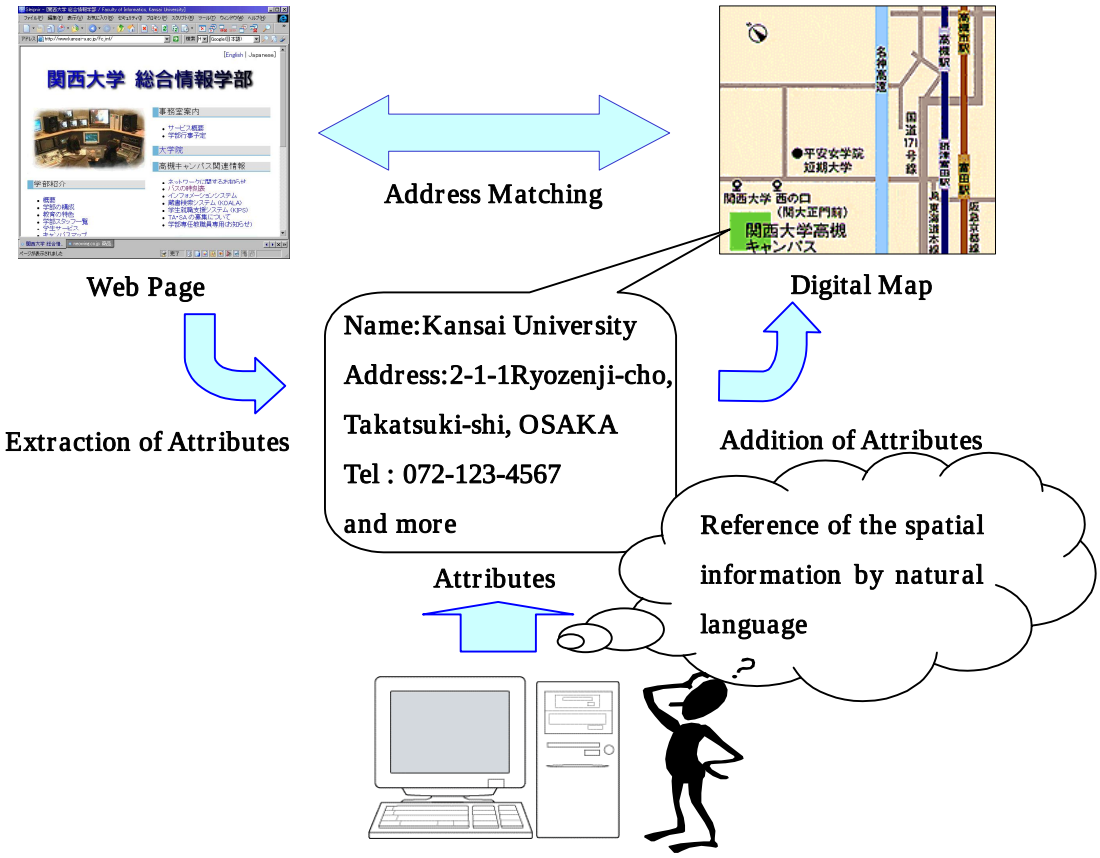

Fig.1 Concept of the Proposed System

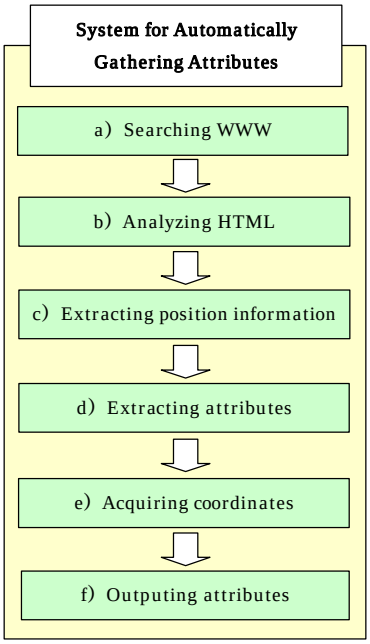

Fig.2 Flowchart of the Subsystem for Automatically Gathering Attributes

3. Details of the System

This section provides details of the two systems we developed for the retrieval of spatial information for digital maps. One system automatically gathers attributes associated with digital maps from the World Wide Web (WWW). Another system provides spatial information search capability based on keywords derived from natural language.

(1) Subsystem for Automatically Gathering Attributes

This system realizes automatic extraction of attribute information on digital map by searching the WWW. This system consists of 6 processes: 1) searching the WWW; 2) analyzing HTML; 3) extracting positional information; 4) extracting attributes; 5) acquiring coordinates; and 6) outputting attributes. This system's flowchart is as shown in **Fig.2**. Details of each function will be explained below.

a) Searching WWW

In this process, search will be performed on the WWW, and Web pages will be collected by tracing link information on the Web pages. Process flow is as

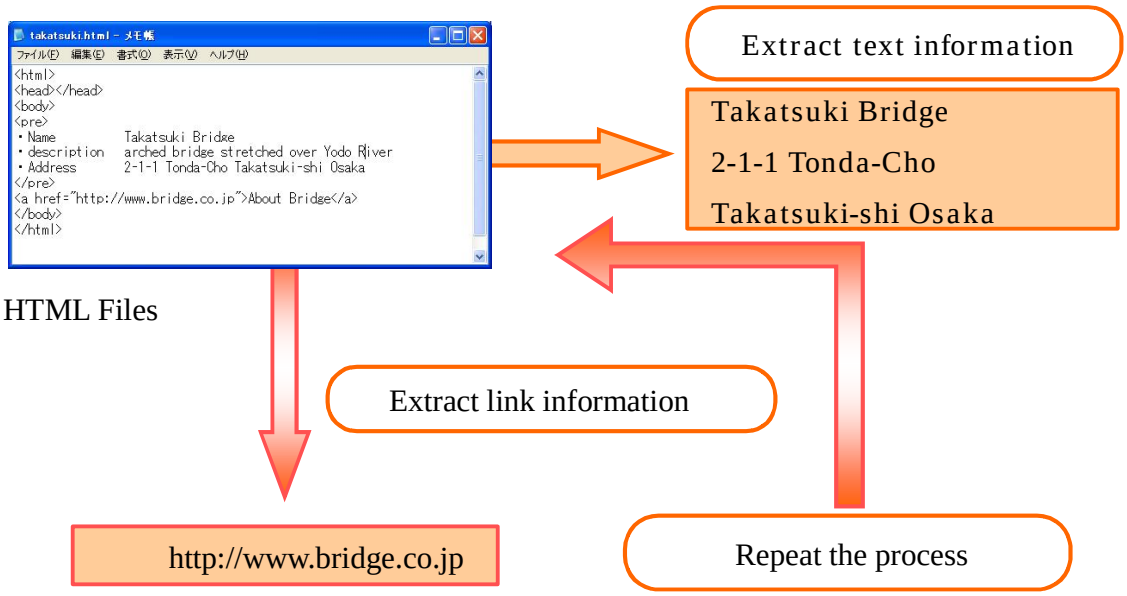

Fig.3 Flow for Searching WWW

shown in **Fig.3**. Files in HTML form are acquired, but EXE files and PDF files are ignored for the following reasons:

- There is a low possibility of such files having address information
- Analysis requires a long time
- Much of the information is not analyzable

HTML files whose URL is a relative path are not acquired, because following links that are relative addresses can lead to repeated searches within the same site. The flow for automatic collection is as follows:

(1) Get a URL from a database
(2) Connect to the Web page of the URL
(3) Acquire the file of the connected Web page
(4) Extract link information from the acquired Web page file
(5) Save the link information in a database
(6) Repeat the process from step (1)

In this process, we must determine the Web page to be used as a reference point. Selecting the reference point page affects the results of information gathering, so it is very important. In our research, we use the following criteria for selecting the Web page to be used as a reference point.

- The page should have abundant link information
- The links of interest should be absolute paths
- There should be spatial information
- There should be reference to a particular theme, such as “sightseeing” or “gourmet” for example
- The page is in Japanese

b) Analyzing HTML

The Web page gathered in the last process is analyzed with HTML document parsing software. In the analysis, URLs are acquired from link tags in the page. Furthermore, text data, except for tag information and image information, is acquired. The URLs and text data are saved in a database.

c) Extracting Positional Information

In this process, address information is extracted from text data acquired in the last process. The extraction of positional information is realized by morphological analysis. In our research, the Java "Sen" morphological analysis system is used. When morphological analysis determines “area” as the part of speech in the processed text, then positional information is extracted (**Fig.4**). In this system, positional information is extracted only when there is one positional information for one Web page.

However, morphological analysis may extract this positional information in an imperfect form. For example, incomplete address information such as "Osaka" or "Takatsuki-shi" cannot be converted to exact coordinate information, so they cannot be used by this system. Rather, perfect address information is required, such as "2-1-1 Ryozenji-cho, Takatsuki-shi, OSAKA.” Such perfect address information can then be processed using pattern rules to locate the parts of speech in a text line. However, there is limitation in extraction of positional information by using pattern rules with morphological analysis, and there are cases when incorrect positional information is acquired. In order to

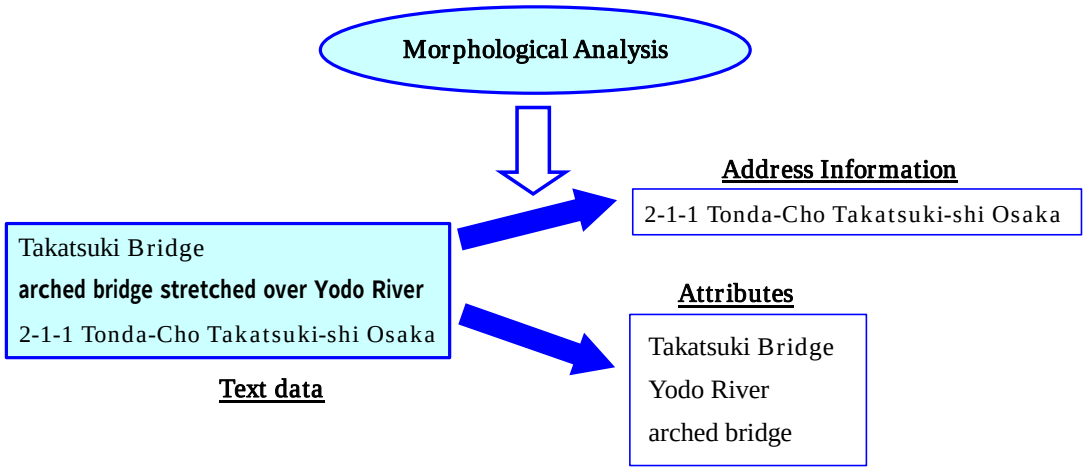

Fig.4 Extracting Address Information and Attributes by Morphological Analysis

solve this problem, address database that contains correct address will be used. In this research, “address postal code download service” offered by Japan Post will be used to produce address database. By checking if the positional information extracted by using morphological analysis exists in the address database, incorrect address can be deleted. With this technique, the accuracy of positional information acquisition can be improved.

d) Extracting Attributes

In this process, attributes are extracted using morphological analysis similar to the acquisition of positional information in the last process. Here, again, we use the Java "Sen" morphological analysis system to locate the appropriate part of speech and extract "nouns," "adjectives" and "verbs." By acquiring information in this way, natural-language words, such as "beautiful," can be used as attributes in the digital map (**Fig.4**). The acquired attributes are saved in a database.

e) Acquiring Coordinates

In this process, coordinate information is determined from the address information acquired by the process described in Section c). Determination of coordinate information is realized through address matching (**Fig.5**) using a Comma Separated Value (CSV) address matching service[8] developed by the Center for Spatial Information Science at the University of Tokyo. By using this service, the address information is converted to coordinates, such as latitude and longitude, making it possible to supply attribute links to the object on a digital map (**Fig.6**).

f) Outputting Attributes

In this process, the URL, positional information,

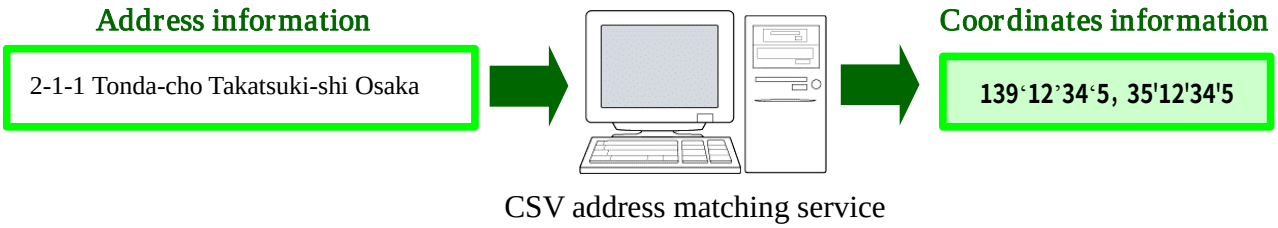

Fig.5 Acquiring Coordinates by Using Address Matching

Enable to link between feature of digital map and attributes from coordinates

Coordinates	Address	URL	Attributes
139'12'34'5 , 35'12'34'5	2-1-1 Tonda-Cho Takatsuki-shi Osaka	http://www. bridge.com	Takatsuki Bridge arched bridge stretched

Fig.6 Linking Land Objects and Attribute Information on Digital Map by Using Coordinate Information

coordinate information and attributes collected by each of the previous processes are outputted as a file in XML form using XML-DOM technology. As XML data, the information can be read and modified easily. **Fig.7** shows an example of collected attribute information output in XML form. The acquired information is managed by a database. For our project, "MySQL" is used as the database system. MySQL, an open source SQL system, is capable of constructing and managing relational databases and the system can search large amounts of data at high speed.

（２）Sub System for Retrieving Spatial Information with Natural Language

This system is used to retrieve spatial information using natural-language keywords. There are five processes: 1) entering a keyword; 2) selecting a search area; 3) creating related word and synonym references; 4) creating object references; and 5) outputting the reference results. **Fig.8** shows a flowchart for the proposed system. The details of each process are described below.

a) Entering a Keyword

In this process, a keyword in natural language, such as "pleasant" or "healing," is provided to the system as input.

b) Selecting a Search Area

In this process, the geographical area the user wants to search is selected. This is done by selecting a digital map file saved beforehand to an arbitrary directory.

c) Searching some Object

In this process, objects relating to a keyword are located on a digital map. Based on the results of the preceding processes, all words and synonyms relating to a natural-language keyword are compared against attributes acquired from objects on the digital map by the "system for automatically gathering attributes." In this way, the system can determine references to those objects on a digital map that are appropriate to a natural-language keyword supplied by a user.

d) Outputting Reference Results

In this process, the reference results produced by the previous processes are displayed on the digital map. Those objects relating to the reference keyword supplied by the user are shown as icons on the map. By choosing an icon with a mouse, the related Web page can be displayed. **Fig.9** shows an example of object references being displayed on a digital map.

4. Experimental Demonstration

In order to verify the usefulness of the developed system, we performed an experimental demonstration. First of all, we performed "attribute information extraction experiment" to measure the accuracy of extraction of attribute information. Next, we performed "coordinate information acquisition experiment" to verify if the coordinate information can be extracted from the extracted positional information. Next,

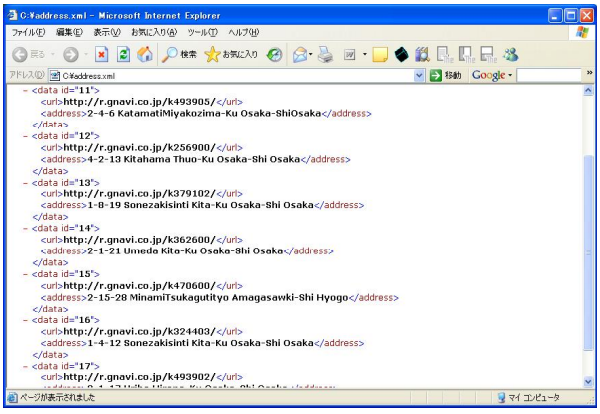

Fig.7 Example of Attribute Information Output in XML

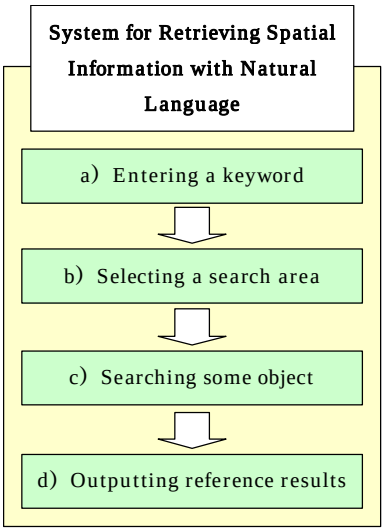

Fig.8 Flow of Spatial Information Retrieval Subsystem

Fig.9 Example of Search Result

"spatial information search experiment" was performed in order to measure accuracy of spatial information search by using natural language. And finally, "user satisfaction level investigative experiment" was carried out in order to investigate the satisfaction level of users who performed spatial information search by using natural language. The details of each experiment are explained below.

(1) Attribute Information Extraction Experiment

a) Method of Experimental

In this experiment, attribute information was gathered by using the automatic attribute information gathering system. We have selected "Gourmet Navigator (http://www.gnavi.co.jp)", "Rakuten Travel (http://www.mytrip.net)", and "web an (http://weban.engokai.co.jp)" for reference point Web pages. The reason for choosing these Web pages as reference point is that they contain abundant amount of address information and link information. In this experiment, link search time was set to be 3 hours, and search was performed on the WWW. After the search, number of address information that was inside the gathered Web pages was measured.

b) Result of Experiment

As a result of the measurement, when "Gourmet Navigator" was set as reference point, there was approximately 10 thousand Web pages that was gathered by automatic WWW search. The system was able to extract address information from 1,717 Web pages. As a result of verifying Web pages that contains address information, there were approximately 1,800 pages that maintain address information. The system's experimental result, including the results of performing search by using other Web pages as reference point, is shown in **Table 1**.

Table 1 Result of Experiment for Attribute Information Extraction

	Gourmet-Navi	Travel info	web an
Number of gathered Web pages	10,025	12,890	15,130
Number of pages that system can extract address from	1,717	1,115	1,457
Number of pages that have address	1,804	1,478	1,750
Accuracy of extracting address	95.2%	75.4%	83.3%

Table 2 Result of Experiment for Coordinate Information Acquisition

Coordinate Accuracy	Number	Ratio
Block number level	3,269	76.2%
City district level	746	17.4%
Acquisition not possible	274	6.4%
Total	4,289	100.0%

c) Consideration

From the experimental result, we were able to acquire address information with the accuracy of 95.4% with "Gourmet Navigation", 74.3% with "Rakuten Travel", and 83.3% with "web an". In this system, because both attribute information and address information are acquired and saved in database, accuracy of positional information acquisition will be the same value as that of attribute information acquisition. Hence, from this experiment, it became evident that attribute information can be acquired with an average accuracy of 84.3%. By this experimental result, it was proven that this system is effective for automatic production of attribute information of digital map data.

(2) Coordinate Information Acquisition Experiment

a) Method of Experiment

Acquisition of coordinate information was performed by using CSV address matching service. In CSV address matching service, accuracy of the coordinates is at a level where the result can be outputted in terms of block number, or city district. Address matching was performed on all 4,289 address information that was acquired in attribute information acquisition experiment.

b) Result of Experimental

As shown in **Table 2**, after address matching, 3269 out of 4289 address information could be transformed into coordinate information. Reversely, we were not able to acquire coordinate information from 274 address information.

c) Consideration

As a result of the experiment, coordinate information was acquired with accuracy within block number from 76.2 % of address information by using address matching. For the reason that accurate link to electronic map would be possible with coordinate information with accuracy within block number, we can speculate that majority of the attribute information are capable of being applied in electronic map. However, performing a link with accuracy within city district is impossible. As a result of the experiment, because 23.8 % of the attribute information cannot be linked with electronic map, we can say that, for the near future, there is a need to increase the accuracy of natural language processing and link much more attribute information.

(3) Spatial Information Search Experiment

a) Method of Experiment

In this experiment, comparison was made between this system and the existing system. In order to verify the effectiveness of the system, comparison was performed with 8 headings of natural language. In order to perform search with the same condition as the system, free word search was used for MapFan, and keyword search was used for Mapion. In addition, because the reliability of the search result is not guaranteed, restaurant portal site "Gourmet Navigation" was used to confirm the relationship between the object name and the search keyword. In this research, comparison was made between Mapion, MapFan, and the system by using "ambiguous natural language" and "words that indicates location".

For the search by using ambiguous natural language as keyword, frequently used ambiguous words, or words that cannot specify objects, will be defined as "ambiguous natural language". The interpretation of "ambiguous natural language" varies depending on one's taste and it is difficult to find a single definition. For example, words such as "healing", "pleasant", "relaxing", "suddenly", and "small barriers" are ambiguous, and they have different way of being interpreted depending on the person. However, people usually decide upon a specific land object after thinking upon these ambiguous words.

In the search by using words that indicate location as keyword, because the words are habitually used, they are less ambiguous, and words that indirectly indicate location are defined as "words that indicate location". Although location cannot be directly specified by "words that indicate location", they are words that can narrow the locations down to a certain degree. For example, although we cannot specify "elementary school", "amusement park", and "hot spring" into 1 location, we can raise a specific land object as a candidate. In the experiment, land objects are retrieved from "words that indicate location", and land objects' extraction result is verified.

b) Result of Experiment

Retrieval by "ambiguous natural language" acquired a result as shown in **Table 3**. For keywords "curing", as Mapion and MapFan were able to acquire 1 and 2 results, respectively, this system was able to acquire 25 results. For the rest of the 3 keywords, the system also acquired a better result, similar to above.

For retrieval by "words that indicate location", a

result as shown in **Table 4** was acquired. For the keyword "hot spring", as Mapion and MapFan both acquired 51 results, this system was able to acquire 21 results. The reverse result was acquired for the keyword "book store".

c) Consideration

In this research, for the retrieval of land information by natural language, by splitting the keywords into "ambiguous natural language" and "words that indicate location", the usefulness of this research was verified. For "ambiguous natural language", the system was able to retrieve more land objects than the existing system. However, for the "words that indicate location", the existing system was able to retrieve more land objects than this system.

The reason that the system was able to retrieve more land objects with "ambiguous natural language" than "words that indicate location", in the existing system is that words that can easily specify land object are categorized, and when land object that is suitable for the category is retrieved, the land object is expressed as a retrieval result. However, words that are not categorized are expressed as a retrieval result only if the word is identical to the land object name or if the word exists in annotation of the land object. For this reason, in this system, we were able to extract more land object names from "ambiguous natural language".

(4) User Satisfaction Level Experiment

a) Method of Experiment

In this experiment, spatial information retrieval with natural language was performed by using the developed system. Investigation of satisfaction level of users was carried out by using the survey result that was obtained after the experiment.

b) Result of Experiment

As a result of the experiment, the survey result as shown in **Table 5** was acquired. The average value of the satisfaction level was 3.77. In the opinions that was gathering during the survey, there were positive comments such as "Very fun to search because I can find places that are related to the words that came up to my mind", "I was able to find new spots that I didn't know of previously". The negatives comments were "Cannot tell whether the location related to the word I had

Table 3 Retrieval Result for "Ambiguous Natural Language"

	Barrier free	Pleasant	Healing	Intellectual
Proposed System	11	41	25	4
Mapion	0	0	1	0
MapFan	0	6	2	0

Table 4 Retrieval Result for "Words that Indicate Location"

	Elementary School	Amusement Park	Hot Spring	Bookstore
Proposed System	30	6	21	8
Mapion	46	8	51	1
MapFan	0	0	51	3

Table 5 Survey Result of User Satisfaction Level

Satisfaction Level	Number	Ratio
5	8	20.0%
4	11	33.3%
3	7	23.3%
2	3	13.3%
1	2	10.0%

imagined is actually being searched", "Too few search results", and "Too many search results".

c) Consideration

For the reason that the satisfaction level exceeded the average value, we can speculate that there was some effect of the search by using the system. On the other hand, because there were users with low satisfaction level, we believe that there is a requirement to undertake development of a system that considers the needs of those users in the future.

In the opinions gathered in the survey, there were examinees that had fresh impression on the search by natural language, and 19 out of 30 responses were positive opinions. We can say that spatial information processing by natural language has been evaluated.

5. Conclusion

In this research, by automatically searching the WWW, development of a system for using information on the WWW as attribute information of digital map has been achieved. By using the automatic attribute information gathering system and spatial information search system that were developed in this research, the following were realized:

- Acquisition of attribute information by automatic search on the WWW.
- Extraction of address information and attribute information from HTML by the use of morphological analysis.
- Maintenance of attribute information by XML.
- Spatial information search by natural language.

In addition, the validity of the system was verified by an experiment.

From the above, we can speculate that the cost and effort for producing attribute information of digital map can be reduced by this research. Furthermore, we can speculate that by allowing digital map to contain detailed attribute information, there can be various uses to the construction field. For example, attribute such as age of a building can be a good use in disaster prevention field; attribute of number of floors and outward appearance can be used in city planning; and attribute of interior facility can be used in barrier free interior design.

Gratitude: This work was supported by "Open Research Center" Project for Private Universities:Matching Fund Subsidy from MEXT(Ministry of Education,Culture,Sports,Science and Technology),2003-2007.

Reference

1) Krzanowski, R., Raper, J. : Spatial Evolutionary Modeling, Oxford University Press, 2001.4.

2) Plewe, B. : GIS Online, Thomson Learning, 1997.8.

3) Ministry of Internal Affairs and Communications : 2004 Version Information Communication White Paper, Gyousei, 2004.7. (in Japanese)

4) Sagara, T., Arikawa, M., Sakauchi, M. : Spatial Information Extraction System Using Geo-Referenced Information, Transactions of Information Processing Society of Japan:DATABASE, Vol.41, No.SIG6(TOD7), pp.69-80, 2000.10. (in Japanese)

5) Nakajima, T., Otsubo, S., Yamana, D., Tomimatsu, A. : Implementation of a Map Search System Using Web Service, Papers and Proceedings of the Geographic Information Systems Association, Vol.12, pp.383-386, 2003.9. (in Japanese)

6) Saito, M., Tanaka, F., Kanaj, S., Kishinami, T. : Geographic Information Query System Using Semantic Web, Papers and Proceedings of the Geographic Information Systems Association, Vol.11, pp.293-296, 2002.9. (in Japanese)

7) Kubo, N., Imura, I., Iida, G., Hirai, M., Otomo, S. : Development of GIS Model System that Integrated Spatial Information and Time Series Information, APA, Association of Precise Survey & Applied Technology, No.75-8, pp.67-74, 2000.3. (in Japanese)

8) Sagara, T., Arikawa, M. : Dispersion Address Matching Service Suited for Address System of Japan, Papers and Proceedings of the Geographic Information Systems Association, Vol.9, pp.183-186, 2000.9. (in Japanese)