RESEARCH FOR EXTRACTING ATTR IBUTES OF BUILDINGS ON DIGITAL MAP FROM WEB RESOURCE USING GENETIC ALGORITHM

Kantaro Monobe [1], Shigenori Tanaka [2], Hitoshi Furuta [3], Yuichi Kato [4], and Hiroshige Nonaka [5]

ABSTRACT

A service that utilizes spatial information is recently increasing with the development of information processing technology. This type of service can now be easily used by everyone through the utilize of personal computer or mobile phone. The problem is that the foundation of the spatial information service, the digital map data, lacks attribute information and is not updated very often. The problem can be blamed on the great deal of money and labor required to create, manage, and maintain digital map data. Therefore, the need for development of a system that can automatically gather attribute information is inevitable. The purpose of this research is to develop a system that can produce digital map's attribute information by gathering variety of information from the World Wide Web automatically. In this research, we aim to extract of name of buildings, which is currently difficult to achieve, by using natural language processing and genetic algorithm.

KEY WORDS

gis, web resources, attributes of digital map, genetic algorithm.

INTRODUCTION

In recent years, the importance of spatial information has increased remarkably, with the development of information technology. Owing to the spread of GPS (Global Positioning System) and improvement in its accuracy, anyone can now easily obtain positional information, and spatial information is becoming one of the necessities of our lives. Services provided by GIS (Geographic Information System), a system that handles spatial information,

[1] Ph.D Candidate, Master of Informatics, Graduate School of Informatics, Kansai University, 2-1-1 Ryozenji-cho, Takatsuki-shi, OSAKA, Japan 569-1095, Phone +81 72/690-2153, FAX +81 72/690-2491, k_monobe@kb.kutc.kansai-u.ac.jp

[2] Professor, Dr. Eng., Faculty of Informatics, Kansai University, 2-1-1 Ryozenji-cho, Takatsuki-shi, OSAKA, Japan, 569-1095, Phone +81 72/690-2153, FAX +81 72/690-2491, tanaka@res.kutc.kansai-u.ac.jp

[3] Professor, Dr. Eng., Faculty of Informatics, Kansai University, 2-1-1 Ryozenji-cho, Takatsuki-shi, OSAKA, Japan, 569-1095, Phone +81 72/690-2153, FAX +81 72/690-2491, furuta@res.kutc.kansai-u.ac.jp

[4] Master's Course Student, Bachelor of Informatics, Graduate School of Informatics, Kansai University, 2-1-1 Ryozenji-cho, Takatsuki-shi, OSAKA, Japan, 569-1095, Phone +81 72/690-2153, FAX +81 72/690-2491, babgj308@jttk.zaq.ne.jp

[5] Master's Course Student, Bachelor of Informatics, Graduate School of Informatics, Kansai University, 2-1-1 Ryozenji-cho, Takatsuki-shi, OSAKA, Japan, 569-1095, Phone +81 72/690-2153, FAX +81 72/690-2491, fa12054@kb.kutc.kansai-u.ac.jp

is increasing. People can use these services through their PC and mobile phone. In order to further develop these services, it is important to maintain information used by GIS.

In GIS, data is organized by the concept of feature. Feature is composed by geometric information and attributes. Thanks to the development of computer graphics technology, geometrical information can now be represented 3 dimensionally. Digital map extremely similar to the real world is being made by the 3D technology. However, we still have a long way to improve the maintenance of attribute information. For example, numerical map from the Publication of Geographical Survey Institute and digital map provided by the Web usually maintain only minimal attribute information such as name of buildings, address, and phone number.

Presently, with insufficient attribute information, sufficient information cannot be offered to the expanding demands for services of spatial information in the future. For spatial information, its information updates are constantly in demand. In order to execute spatial information service accurately, spatial information equivalent to that of the real world is required to be maintained at all times (Krzanowski and Raper 2001). Frequent updates and maintenance of feature's attribute information are indispensable in achieving this. Under the present circumstances, attribute information maintenance requires considerable amount of money and labor (Plewe 1997). For this reason, we are looking to develop a new system that can easily maintain attribute information.

The existing researches try to utilize the information on the web for the benefit of spatial information maintenance. One of this research tries to extract existing geo-reference information from the Web, convert it to positional coordinates, and puts spatial information service into effective use (Sagara et al. 2000). We are developing spatial information extracting system and spatial information search engine as actual systems. Furthermore, we have proposed to embed spatial tag that uses XML representation, and are doing evaluation on the spatial tag's usefulness. Also, we are building map search system using web services (Nakajima et al. 2003). This service utilizes SOAP messaging and enables distribution of map information. Another research on geographic information search system that uses Semantic Web is being conducted (Saito et al. 2002). However, these researches have yet to perform automatic extraction of attribute information. Also, websites that allow automatic extraction are very limited. In order to solve these problems, it is necessary to analyze the meaning of the web page using natural language processing.

Although name of buildings, one of attribute information, is essential information just as essential as address for displaying positional information, feature name's maintenance is behind. When general users specify a building, they do not specify it from its direct reference that employs the coordinates of latitude and longitude. They specify a building from its indirect reference that uses name of buildings and address information. For example, when you are telling the location to someone, the name such as "Faculty of Informatics in Kansai University" is employed, not "Latitude 31.41, longitude 135.29."

There are previous researches on maintenance of building's name. One is a research that applies spatial contents on the Web to GIS (Sagara et al. 2003). Another is a research on a system that enables general users to put spatial information on record (Sagara and Arikawa 2001). Because the former extracts the Web's name of buildings on Townpage, the accuracy

of building's name is dependent on the accuracy of Townpage. Real-time information is the advantage of the latter, but it does not lighten the burden of updating work.

So in this research, we are aiming to do research and development on a system that aids production of building's attribute information, by performing automatic search on the World Wide Web. Furthermore, we will conceive a method for extracting names of buildings, by employing natural language process and genetic algorithm. This is because buildings and building's names are particularly important to attribute information.

OUTLINE OF THE RESEARCH

We have focused on the information from the World Wide Web, as data source for digital map's attribute information. In Japan, there are currently 85.9 million web pages (Ministry of Internal Affairs and Communications 2004). By matching positional information included on all these web pages with digital map's features, we believe that we can attach information from the World Wide Web as buildings' attribute information.

We are aiming to do research and development on a system that aids production of digital map's attribute information, by conducting automatic search on the World Wide Web. This system's outline is shown on figure 1.

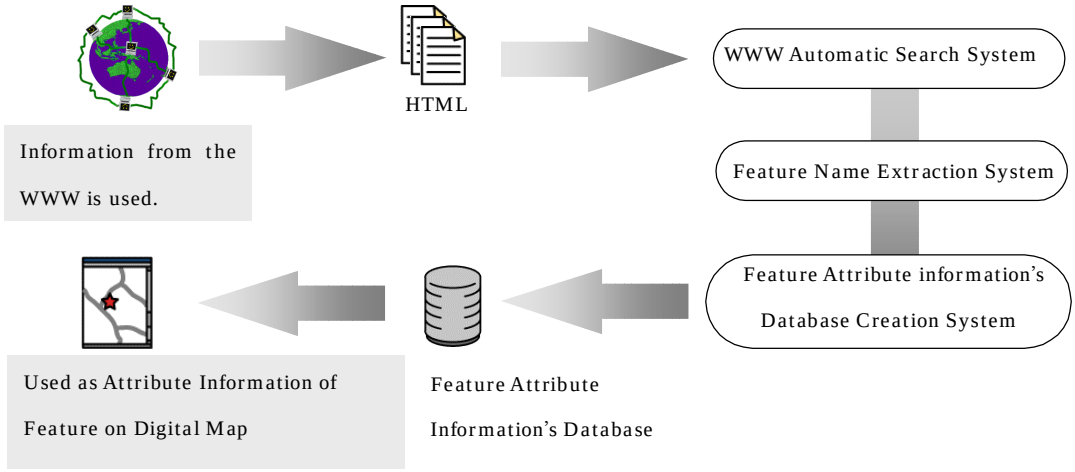

Figure 1: Outline of the Proposed System

This system first extracts information related to buildings from web pages. Then, by using this information, it extracts building's names and classifies features.

This thesis is organized in the following way. Method that performs automatic search on the WWW and extracts address information and its accompanying information is explained in chapter 3. Method for extracting feature names by employing natural language processing is mentioned in chapter 4. Method that outputs acquired attribute information on digital map in usable condition is mentioned in chapter 5. And finally in chapter 6, the fruits that can be obtained from this research and examination on the problems we may encounter in the future are mentioned.

WWW AUTOMATIC SEARCH SYSTEM

In this chapter, we will mention about a method for extracting digital map's address information and its accompanying information on building's account by performing automatic search on the WWW. This system is realized by the following 4 processes: 1) WWW automatic search; 2) HTML analysis; 3) Address information extraction; 4) Extraction of information on building's account. Details on each process are explained below.

WWW AUTOMATIC SEARCH

In this process, a search on the WWW is performed by tracing link information within web pages and gathers web pages. Files in HTML form are acquired. Other files such as EXE files and PDF files are ignored whether they are on web pages' links or not for the following reasons: The chance of EXE file having address information is small; analysis of PDF file's information requires a great deal of time. The procedures of WWW automatic search are listed below.

1) Acquire URL from a database

2) Go to the URL's web page.

3) Acquire the web page's file.

4) Acquire link information within the web page.

5) Save link information on the database.

6) Repeat the procedures from 1 to 5.

In order to perform WWW automatic search, it is necessary to determine in advance, the web page to be used as a reference point. Gathered information will vary significantly depending on the Web page used as a reference point. Therefore, the selection of the web page to be used as a reference point is crucial. In this research, we use the following criteria for selecting the Web page to be used as a reference point.

- The Web page should have abundant link information.
- The links of interest should be absolute paths.
- The Web page should have relevance to place names and locations.
- The Web page should have reference to themes, such as "Sightseeing" and "Gourmet".
- The Web page should be in Japanese.

By the considering the above criteria, WWW automatic search is realized.

ANALYZING HTML

Web pages gathered by the WWW automatic search is analyzed with HTML document parsing software. This is software used for reading HTML document and analyzing it. By

analysis, URL is acquired from link tag in the Web page. Furthermore, web page's text data, excluding tag and image information, is acquired.

EXTRACTING ADDRESS INFORMATION

In this process, address information is extracted from text data acquired by HTML analysis. The extraction of address information is realized by morphological analysis. In our research, Java Morphological Analyzing System called "Sen" is employed. When morphological analysis determines "Area" as part of the speech in the processed text, address information is extracted. In this system, address information is extracted only when the web page has one address.

However, morphological analysis may extract incomplete address information such as "Osaka-fu" or "Takatsuki-shi", which cannot be converted to exact coordinates information, so they are useless. Instead, a complete address information such as "2-1-1 Ryozenji, Takatsuki-shi, Osaka-fu" is required to be extracted. Such complete information can be extracted using pattern rules to locate parts of the speech in a text line. However, there is a limitation on the extraction of address information using pattern rules with morphological analysis, and there are cases when incorrect address information is acquired. To solve this problem, address database containing correct addresses is used. In this research, "District-level Positional Reference Information Download Service" (Ministry of Land, Infrastructure and Transport 2005) offered by the Japan Ministry of Land, Infrastructure, and Transport is used to create address database. By checking if the address information extracted using morphological analysis exists in the address database, incorrect addresses can be deleted. With this technique, the accuracy of positional information acquisition can be improved.

However, information on blocks and lot numbers such as "1－7－3" or "1st block, 7th lot, 3" cannot be extracted from address database. So by using pattern matching, which employs proper expression, to acquire lot information, complete positional information can be extracted. Moreover, special address book is designed for address names found in central part of Kyoto, to extract its positional information. By that, we can improve the precision of address information extraction.

EXTRACTING FEATURE'S ACCOUNT INFORMATION

In this process, building's account information is extracted using morphological analysis in the same way as extracting address information in the last process. We use the Java morphological analyzing system "Sen" to locate appropriate parts of speech and extract "Nouns", "Adjectives", and "Verbs". By acquiring information this way, attribute information, such as "Beautiful" and "Delicious", which did not exist on previous maps, can be obtained to be used on digital maps today. The acquired attributes are saved in a database.

FEATURE NAME EXTRACTION SYSTEM

In this chapter, the method for acquiring building's name from its corresponding address is mentioned. This system uses characteristics related to building's name such as "(1)Feature name exists near the letters of the address in a Web page", and "(2)New conceptual term such as building's name is described as compound word most of the time" [11]. This system is

realized by the following 4 processes: 1)Extraction of information related to the address; 2)Preprocess for morphological analysis; 3)Calculation of morpheme's extraction and its importance level; 4)Extraction of compound word by genetic algorithm. Details on each process are explained below.

ACQUISITION OF ADDRESS RELATED INFORMATION

In this process, address information acquired by the WWW automatic search system is used as a source for creating the following 4 types of address information: (1)Complete address such as "2-1-1 Ryozenji-cho, Takatsuki-shi, Osaka-fu"; (2)Address that has its first part altered such as "2 cho-me 1-1 Ryozenji-cho, Takatsukishi, Osaka-fu"; (3)Address that represents a district such as "Ryozenji-cho, Takatsuki-shi, Osaka-fu"; (4)Address has its municipal name omitted such as "2-1-1 Ryozenji-cho, Takatsuki-shi". These patterns are used to increase the acquirable web pages. By querying these patterns, web page search via GoogleAPI is conducted. Address character row's 50 preceding letters and 50 following letters are gathered from the acquired web page. Hereafter, this character row will be addressed as "Address related character row".

PREPROCESS FOR MORPHOLOGICAL ANALYSIS

Preprocessing is conducted in order to prevent extraction of character row irrelevant to building's name. Address, phone number, postal code with no relation to building's name have been excluded from "Address related character row".

CALCULATION OF MORPHEME'S EXTRACTION AND ITS IMPORTANCE LEVEL

In this process, morphological analysis is conducted on "Address related character row", then morpheme's importance level is calculated using TF*DF method, which is a combination of "TF method" and "DF method". In morphological analysis, nouns, verbs, and unknown words with high potential of being compound word (building's name) are used. In TF*DF method, when many morphemes are included in one web page and when many morphemes are included in many web pages, importance level of morphemes' is judged as high.

EXTRACTION OF COMPOUND WORDS VIA GENETIC ALGORITHM

In this process, compound words are extracted from morphemes gathered by TF*DF method. Genetic algorithm is used for the reason that extremely large number of morpheme combinations appear when extracting compound words. Number of morpheme combinations is represented in formula (1).

$$\sum_{k=0}^{k=n-1} {}_n P_{n-k} \qquad (1)$$

In this formula, "n" represents the number of morphemes. The gene shows whether or not it includes morpheme. If the gene includes morpheme, it is 1, and if it does not include morpheme, it is 0. Crossing-over uses the method that reserves 2 arrangements: One, with the highest estimated value, and another, with the second highest estimated value. As for mutation, a method that randomly crosses over 1 point is used.

The following 3 characteristics are employed when extracting compound words: (1)Frequencies of each group of morphemes, which compose compound word, appearing in the text will all be the same. As a result, there will be less distribution of morpheme's TF*DF value; (2)The average TF*DF value of letters that form compound word increases depending on the letters' importance in the text; (3)Since compound word is not a letter, the number of words that form a compound word increases. Characteristic (1) uses the reciprocal of distributed TF*DF value, (2) uses the average of TF*DF value, and (3) uses the number of letters in a compound word. Characteristic (1), (2), and (3) exists only as much as the number of the combination of genetic algorithm. Because the acquisition of complementary set's average and standard deviation is difficult due to the large number of combinations, when searching for characteristic (1), (2), and (3)'s standardized data, randomly extracted sample data is used. Sample's combination will be represented as $C_n = \{c_1, c_2 ... c_n\}$. For each components, the average of TF*DF value is represented as $CC_n = \{cc_1, cc_2 ... cc_n\}$; distribution as $S_n = \{s_1, s_2 ... s_n\}$; and number of words as $L_n = \{l_1, l_2 ... l_n\}$. C_n's component c_1 is represented as $C_1 = \{a_1, a_2 ... a_r\}$, which indicates a set of morphemes used when combining within the sample. Each of its components, a_r, represents TF*DF value of morphemes gathered from each Web page. And the combination that will be inquired is represented as $Q_m = \{q_1, q_2 ... q_m\}$. Formula for estimation, Eva, is indicated in formula (2).

$$Eva = \frac{H_{\bar{Q}} - Min}{H_{S_Q} - Min} \bullet \log(H_{L_Q} - Min + e) \quad (2)$$

$$\text{Assuming } Min = \min(H_{\bar{Q}} \mid H_{S_Q} \mid H_{L_Q}) - 1$$

Formula for estimation, Min, represents the minimum amount of data within standardized data. H_Q represents standardized data of Q, H_{S_Q} represents standardized data of S_Q, and H_{L_Q} represents standardized data of L_Q. Standardized data H_Q, H_{S_Q}, and H_{L_Q} is obtained using the combinations of the sample and applying them to formula (3)-(5).

$$H_{\bar{Q}} = \frac{(\overline{Q} - \overline{CC})}{v_{c_n}} \quad (3), \quad H_{L_Q} = \frac{(L_Q - \overline{L_n})}{v_{L_n}} \quad (4), \quad H_{S_Q} = \frac{(S_Q - \overline{S_n})}{v_{s_n}} \quad (5)$$

In the above formulas, S_Q represents distributed Q, and L_Q represents the length of Q. v represents the standard deviation for each.

EXAMPLE OF FEATURE NAME ACQUISITION PROCESS

As this system's example of execution, "1 cho-me, 14-15 Shinmachi, Nishi-ku, Osaka-shi, Osaka-fu" is used. In this example, "Osaka Koseinenkin Kaikan" is the building's name. In this example of execution, since there are 20 morphemes, there are 661 quadrillion combination problems. The number of sample data is calculated assuming that reliability is 95% and maximum margin of error is 1%, and 9641 sample data, the answer, is adopted. When "1-14-15 Shinmachi, Nishi-ku, Osaka-shi, Osaka-fu" is inputted, morphemes and their TF*DF value are calculated from the row of letters near the address, as shown on table 1.

Compound words were extracted by means of genetic algorithm in the order shown on table 2.

In this example of execution, since the TF*DF value of "Kaikan", "Kosei", "Osaka", and "Nenkin" are all the same, there will be more distribution. Because the average TF*DF value is high, it is conceived that compound words were extracted when "Osaka Koseinenkin Kaikan" was inputted.

Table 1: Morpheme and TF*DF Value

Morpheme	Part of Speech	TFDF Value
—	unknown word	4.12862
•	symbol, general	1.85315
Kaikan	noun, general	1.45042
Kosei	noun, general	1.45042
Osaka	noun, proper noun, region, general	1.45042
Nenkin	noun, general	1.45042
Ward	noun, suffix, region	1.05554
City	noun, suffix, region	1.05554
Hotel	noun, general	1.05554
Prefecture	noun, suffix, region	1.05554
Wellcity	unknown word	1.04858
Minute	noun, suffix, general	0.65101
Station	noun, suffix, region	0.65101
Ru	unknown word	0.55785
Four	noun, number	0.54766
Tsu	unknown word	0.54766
Bridge	noun, suffix, general	0.54766
Room number	noun, suffix, general	0.54766
Yoshimoto	noun, proper noun, organization	0.51329
2	unknown word	0.44628

Table 2: Extraction of Compound Word by Genetic Algorithm

Combination of morphemes	Evaluation Value
Kaikan	3.4324
Nenkin Kaikan	4.5256
Osaka Kosei Nenkin	5.6188
Osaka Kosei Nenkin Kaikan	6.7119

FEATURE ATTRIBUTE INFORMATION'S DATABASE CREATION SYSTEM

A method for setting up gathered attribute information in usable condition and outputting it on digital map is mentioned in this chapter. This system is realized by the following 2 process: 1)Coordinates information acquisition process; 2)Output process of attribute information. Details on the 2 processes are described below.

COORDINATES INFORMATION ACQUISITION PROCESS

In this process, coordinates information is determined from address information, which is acquired by WWW automatic search system. Determination of coordinates information is realized by address matching, or geocoding. CVS (Comma Separated Value) Geocoding Service (Sagara and Arikawa 2000) offered by Center for Spatial Information Science at the University of Tokyo is used when performing address geocoding. By using this service, address information is converted to coordinates, such as latitude and longitude, making it possible to supply attribute links to the object on digital map.

OUTPUT PROCESS OF ATTRIBUTE INFORMATION

In this process, attribute information such as building's names, classified building's names, URL, address information, coordinates information, and building's account information, gathered in the last process, are outputted as a file in XML format using XML-DOM technology. The reason why attribute information is outputted as a file in XML format is because processing in the program becomes easier, and reading and processing by GIS can be realized. Gathered information will be managed by a database. "MySQL" is employed in 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.

CONCLUSION

In this research, development of a system for using information on the WWW as attribute information of digital map has been achieved, by performing automatic search on the WWW. By using this system, the following were realized:

- Acquisition of attribute information by WWW automatic search.
- Extraction of address information from HTML by morphological analysis.
- Extraction of building's names using genetic algorithm.
- Maintenance of building's attribute information by XML.

From this research, we can speculate that the cost and effort of producing digital map's attribute information can be reduced. Furthermore, since the attribute information contain building's names, we can speculate that there can be various uses in spatial information service. Also, because attribute information is now able to carry natural languages, such as "Beautiful" or "Delicious", which did not exist in attribute information previously, we can speculate that this system is applicable to building search service based on natural language.

However, since this research uses web information as data source for attribute information, its accuracy and reliability is dependent on the web information itself, which is a problem. By conceiving a new method to solve this problem, we can consider a possibility of further improvement on the accuracy.

GRATITUDE

This work was supported by the subsidy from MEXT (Ministry of Education, Culture, Sports, Science, and Technology)(2003-2007), in the category, "Open Research Center" Project, which falls under MEXT's subsidy system titled "Promotional Project for Advancement of Academic Researches at Private Universities".

REFERENCES

Ministry of Land, Infrastructure and Transport (2005). "District level Positional Reference Information Download Service.", < http://nlftp.mlit.go.jp/isj/> (in Japanese)

Koyama, T. (2001). " An Approach to Structural Analysis of Japanese Composite Terms by Supplying Verbs." NII journal, Vol.2, 39-44 (in Japanese)

Krzanowski, R. and Raper, J. (2001). "Spatial Evolutionary Modeling." Oxford University Press

Ministry of Internal Affairs and Communications (2004). "2004 WHITE PAPER Information and Communications in Japan." Gyousei. (in Japanese)

Nakajima, T., Otsubo, S., Yamana, D., Tomimatsu, A. (2003). "Implementation of a Map Search System Using Web Service." Papers and Proceedings of the Geographic Information Systems Association, Vol.12, 383-386 (in Japanese)

Plewe, B. (1997). "GIS Online." Thomson Learning.

Sagara, T., Arikawa, M. (2000). "Distributed Address Matching Service Suited for Address System of Japan." Papers and Proceedings of the Geographic Information Systems Association, Vol.9, 183-186 (in Japanese)

Sagara, T., Arikawa, M., Sakauchi, M. (2000). "Spatial Information Extraction System Using Geo-Referenced Information." Transactions of Information Processing Society of Japan:DATABASE, Vol.41, No.SIG6(TOD7), 69-80 (in Japanese)

Sagara, T., Nakayama, M., Noaki, K., Sadahiro, Y. (2003). "Developing WEB Yellow Page with Map Interface." Papers and Proceedings of the Geographic Information Systems Association, Vol.12, 323-326 (in Japanese)

Sagara T., Arikawa M. (2001). "E-Mail Based Geographic Information System: Post GIS." Transactions of Information Processing Society of Japan, Vol.2001, No.71, 3-8 (in Japanese)

Saito, M., Tanaka, F., Kanai, S., Kishinami, T. (2002). "Geographic Information Query System Using Semantic Web." Papers and Proceedings of the Geographic Information Systems Association, Vol.11, 293-296 (in Japanese)