Discovery of implicit feature words of place name

Hirokawa, Sachio
Research Institute for Information Technology, Kyushu University

Nakatoh, Tetsuya
Research Institute for Information Technology, Kyushu University

Suzuki, Takahiro
Research Institute for Information Technology, Kyushu University

Nakae, Hiroto
Graduate School of Information Science and Electrical Engineering, Kyushu University

出版情報：Proceedings - 2014 IIAI 3rd International Conference on Advanced Applied Informatics, IIAI-AAI 2014, pp.561-566, 2014-01-01. Institute of Electrical and Electronics Engineers Inc.
バージョン：
権利関係：

Discovery of Implicit Feature Words of Place Name

Sachio Hirokawa*, Tetsuya Nakatoh*, Hiroto Nakae†, Takahiro Suzuki*
*Research Institute for Information Technology, Kyushu University,
6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, JAPAN
Email: {hirokawa,nakatoh,suzuki}@cc.kyushu-u.ac.jp
†Graduate School of Information Science and Electrical Engineering, Kyushu University,
6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, JAPAN
Email: nakae.hiroto.068@s.kyushu-u.ac.jp

***Abstract*—Individual opinions and experiences are published in Web as CGM (consumer generated media). A tourism blog which a tourist wrote his experience and impression in a certain area is very helpful information for other tourists. However, a user cannot obtain such precious information without knowing the relation of blog articles and concrete place-names. We paid our attention to the hierarchical structure of place-names. In this paper, we propose the method of connecting related words to the place-name which does not appear explicitly in a blog article paying attention to the hierarchical structure of place-names. From from 45,553 blog articles about the Karatsu area in Saga Prefecture, the potential related words about 78 place-names of Saga Prefecture which have not appeared in the blogs were extracted. 4 subjects evaluated that meaningful related words are obtained in 80% or more of the place-names. However, the direct relationships between the place-name and related words was not able to be guessed easily.**

I. Introduction

In recent years, development of the Internet made publication of information easy. An individual as well as a company and a local government can publish information easily. We pay our attention to sightseeing information in the Internet, and are studying the information retrieval of them.

The tourism information on Web can be classified into three groups. That is, the information that is published by tourist organization in the tourist resort (tourist facilities etc.), and the information that the tour companies offer, and the information in tourism blogs based on personal experiences. Unlike the information by the organizations or by the tour companies, the information in blogs based on the tourist's viewpoint is helpful in many respects. The hidden special features which can be discovered only by actually visiting the tourist resort may appear in them. Those information can be important not only for tourists but also for the tourist boards of local governments and for the tour companies.

However, in a personal blog article, neither a place-name nor a facility name is explicitly written in many cases. Moreover, since the structure of blog articles varies widely, it is not easy to extract and utilize required information in them. Furthermore, if concrete named entities are not known in advance, appropriate search results cannot be expected.

Experiencing new things that is not known in advance and experiencing foreign cultures are important purposes of tourism. However, it is difficult for the internet user to search new thing that is not in their knowledge. When a tourist searches what is in the destination without the detailed knowledge, usually the result is full of information that he/she already knows, and the new information that he/she really wants is rarely acquired.

In order to solve the problems described herein, we have tried to search for the special feature of every place using the hierarchical structure of place-names. The hierarchical structure of place-names is an inclusive relation of the area so that Tenjin is in Chuo-ku, Fukuoka-shi. We expect that the hierarchical structure of place-names enables a user to investigate what kind of area is around the destination, and what kind of resources for tourism are in the area.

It is expected that the famous things and new special features in the area can be discovered by using the hierarchical structure of place-names. Nowadays, a special feature of a narrow area sometimes draws broad attentions. For example, "B class gourmet" (cheap delicious recipes in a local area) attracts large number of peoples all over Japan. Considering such a background, discovery of the special features in narrow areas is useful not only for an Internet user but also for the local government which plans town revitalization.

In this paper, we attempted to use the hierarchical structure of place-names for obtaining special features of Karatsu area in Kyushu, Japan. We experimented for the purpose of obtaining the more peculiar special feature by improving the feature of each prefecture in Kyushu obtained by the general method using the hierarchical structure of place-names. We will report the experiment outline, the obtained result, and the consideration of the result.

II. Related Work

The handling of the tourism information using computers has been studied for many years. Martinez [1] proposed the method for asking by natural language to the knowledge arranged as ontology, and applied it to search of the tourism information. Esparcia [2] created the tool which extracts and incorporates the information on external social networks, in order to reduce the cost of management and maintenance of the tourism recommendation system. Hao [3] proposed a framework called Location-Topic model, and extracted and evaluated the knowledge which represents the location from the travelogues.

Research of Named Entity is important in the viewpoint of the knowledge which symbolizes location. Kinjo [4] used the tag pattern of HTML peculiar to them in order to extract

NE about the spot and event of tourism from Web. Nanbu [5] applied machine learning to the surface pattern about NE, and extracted the tourist spots and the souvenirs from the tourism blogs. Nakatoh [6], [7] extracted foods and tourism spots peculiar to every place using the deviation of the appearance of the noun by the area obtained from a Japanese dependency analysis.

In order to identify the related location from the contents of a blog article, advanced handlings of a place-name is important. Estimation of a blogger's location has been performed in order to extract the information related to the location from a blog article. However, neither an author's profile nor a server's IP address necessarily indicates the location of the contents of a blog article. Fink [8] proposed the method of identifying an author's location from the contents of the blog article. Amitay [9] also performed extraction from blog articles, they treated the ambiguity about the word and name of the same notation as a place-name, and the ambiguity about another location with the same place-name. Toda [10] identified the region of the text characterized by the place-names in a document, and associated the feature words and the place-name of the region using the score. Borges [11] combined a geographical dictionary and geocoding technology with ontology. They have connected the service and the activity which were shown in the Web document to the location using the ontology.

In the existing researches, the target was the location which appears in each blog. In this paper, we pay attention to locations which do not appear explicitly in each blog. The potential locations which are related to each article are extracted by expanding place-names using the hierarchical structure of place-names. We attempt to extract the feature of a location more appropriately by improving the cooccurrence relation of a place-name and related words.

III. Locations Data In The Japanese Postal Code Directory

There are free databases of place-names in Japan, such as the "Supplemental dictionary of place-name for ATOK [1]" or "Gazetteer Of Japan" by Geospatial Information Authority of Japan. However, none of them does not cover the entire place-names in Japan.

The Japanese Postal Code Number data released by the Japan Post Office has the best coverage, and is arranged in hierarchical structure. In this research, we use place-name data in the postal code directory.

We chose Karatsu area (a famous tourist spot) in Saga Prefecture as the research subject in this paper. From the postal code directory, we have extracted place-names in Saga Prefecture and the hierarchy of the places.

Each postal code record is in CSV format. A record consists of ID, reading of the place-name, and the 3 level hierarchical description of the place, that is, prefecture, regional name, and the detailed name of the place. An example of the hierarchical description is like this: "佐賀県, 佐賀市, 今宿町 (Saga Prefecture, Saga City, Imashuku Machi)". However, there are records that have the exceptional "detailed name" fields which contain other than a place-name, such as range of addresses or the description "all other areas than the listed above". We have excluded data which contains any of those descriptions.

Moreover, some of regional name fields and detailed name fields contain two or more elements of place-name such as " X City Y Ward" or "W Town V District". We have separated those descriptions by using the following Kanji characters as delimiters: "都, 道, 府, 県, 市, 郡, 区, 町, 村, ...". They are equivalent of "prefecture, city, ward, town, village etc." in Japanese. For example, from fields "佐賀県, 唐津市, 厳木町牧瀬", 4 level hierarchical description "佐賀県 ,唐津市, 厳木, 牧瀬" is obtained because there is a delimiter 町 in " 厳木町牧瀬".

There are 872 postal code records in the postal code directory data under Saga Prefecture. We have extracted 867 places and the hierarchical relationship between them.

IV. Blog Data

We have collected articles containing the keyword 唐津 (Karatsu) from blog sites. The results was about 47,000 articles written during 2007 to 2013.

We have eliminated duplicated articles by the following procedure. First, we havd divided the articles into sentences. The punctuation mark in Japanese (。) is used as the delimiter of sentences. We got 796,720 raw sentences. Next, we have eliminated sentences which appear twice or more as duplications. It resulted in 667,987 unique sentences. Finally, only the articles containing the unique sentences were restored from the sentences. In our research, we have analyzed those 45,553 original blog articles obtained in this way (Table. I).

TABLE I. Blog Article Statistics

	articles	sentences	words	places
Raw data	46899	796720	114636	603
Without duplication	45553	667987	110057	592

V. The Proposed Method

First, We have built an index for the original blog articles about Karatsu using the morphological analyzer Mecab[2]. The prefix "l:" is attached to the place-names in Saga Prefecture in order to distinguish them from general words. In our analysis, only those words with the prefix "l:" are treated as places. We call the search engine built in this way BSE (the Base Search Engine).

Next, we have built the whole words list $W_i = w_{i.1}, ..., w_{i.n}$ and the list of places $Loc_i = loc_{i.1}, ..., loc_{1.m}$ that appeared in each blog d_i by using BSE. We also built the list of *upper-places* $upper(loc_{i.j}) = \{loc_{i.j.1}, ..., loc_{i.j.k}\}$ for each $loc_{i.j}$ (Here, $loc_{i.j.k}$ is an ancestor of $loc_{i.j}$ in the hierarchical structure). We also call $loc_{i.j}$ is a *lower-place* of $loc_{i.j.x}$. Following to that, we have built an index for all blogs d_i with the whole words and the places including upper-places. We call the search engine built in this way XSE (the eXpanded

[1] http://www.vector.co.jp/soft/win95/writing/se340164.html

[2] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

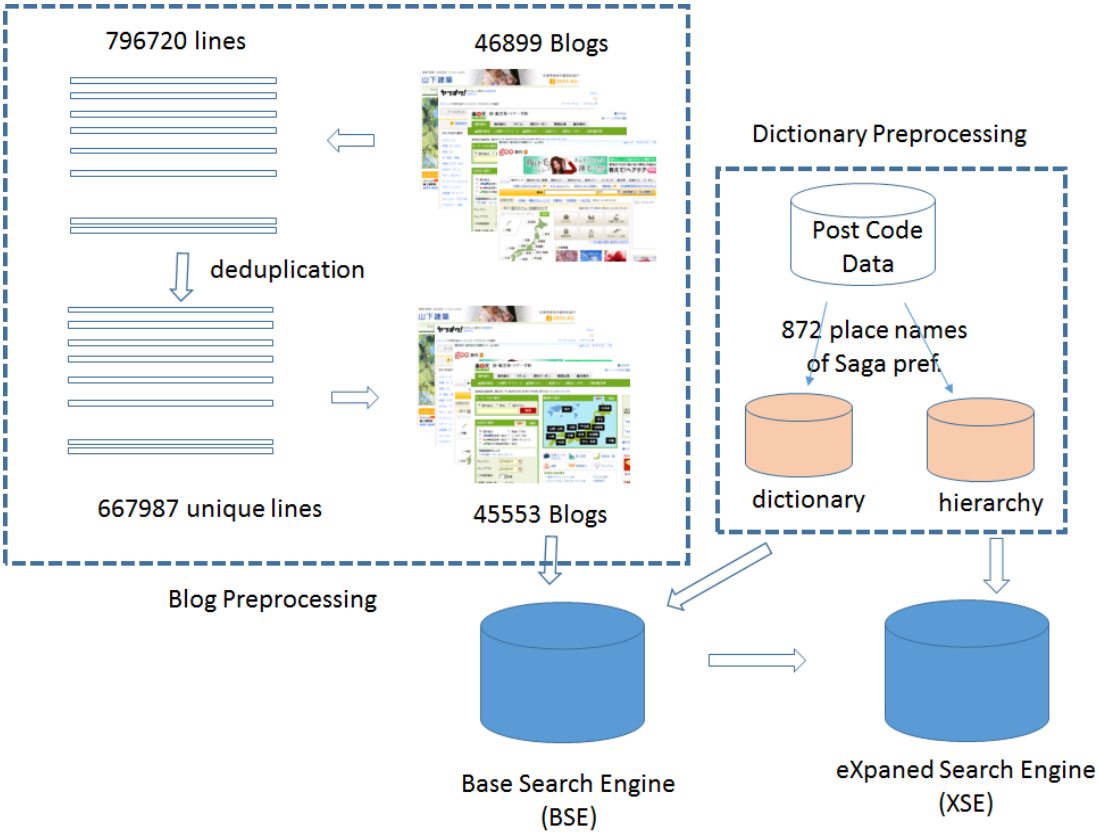

Fig. 1. Construction of BSE and XSE

Search Engine). We also call the procedure for building XSE "expansion of place-names". Fig. 1 shows the outline of our system.

When one searches a place loc which does not appear in blogs by using BSE there will be no hit. On the other hand, when loc is searched by XSE and one of the lower-places of loc appears in a blog, that blog is counted as a hit in XSE search of loc. As a result, the related words of lower-places (loc_1, loc_2, ..., and loc_i) can be extracted as the related words of loc. Related words of places which do not appear in a blog can be discovered in this way. Furthermore, when there is a related word common to loc_1 and loc_2, XSE will extract it as a related word of loc_i ($loc_i$is a sibling of loc_1 and loc_2 in the hierarchical structure) as well as a related word of loc_1 and loc_2.

VI. Feature Words of Unoccurring Location Name

We have conducted an experiment in order to evaluate if significant related words are obtained by the expansion of place-names in XSE described in the previous section.

4 subjects living in Fukuoka Prefecture (Fukuoka is contiguous to Saga) were employed in the experiment. They had not lived in Saga Prefecture. Although they have a certain amount of knowledge about Saga and Karatsu, they do not necessarily know those places in detail.

The following is the detail of the evaluation method. For each place-name loc in Saga Prefecture, we have computede $df(XSE, loc)$, that is the number of the blogs which have hits in an expanded search on XSE. We have compared $df(XSE, loc)$ with $df(BSE, loc)$, the number of the hit-blogs in BSE search. We have extracted 78 locs that have more hits in XSE search than in BSE search. Then, 4 subjects have evaluated whether the related words of the 78 places searched by XSE are appropriate or not.

We have selected related words and related places for each loc from the search result of XSE. That is, the top 20 high scored general words in the standard word score SMART of GETA[3] and top 5 high scored place-names.

Next, the co-occurrence association chart of those 26 words (loc + 5 related places + 20 related general words) for each loc is presented to the subjects as MindMap [12]. The loc is represented in red and the related places are colored in green in the MindMap (Fig. 2).

Each subject judges if (A) he/she can interpret the meaning of related words as groups of words (B) he/she can understand the link between the place-names and related words. Evaluation is perfomed by using five grades, respectively. Finally, the average estime of 4 subjects is taken for each place. The followings are the English translation of the questions.

(A) Can you interpret related 20 related words as groups?

[3]http://geta.ex.nii.ac.jp/geta.html

1) No, not at all.
2) Only a limited part can be grouped.
3) Some parts can be grouped. Others can't be.
4) I can think of a summarizing word that interpret each group.
5) Perfectly.

(B) Can you understand the link between names of places and 20 related words?

1) No, not at all.
2) I can see at least one link.
3) I can understand two or more links between related words and places in green.
4) I can understand links between related words and places in green and in red, respectively.
5) Perfectly.

Fig. 2 is a MindMap showing the cooccurence relation of 10 related words and 5 related places when l:鎮西町 (Chinzei-Chou) is searched by XSE. Even when a word 鎮西町 does not appear in the original blogs, lower-places may appear in some of them. As a result, the upper-place (鎮西町 displayed in red) is linked to the related words of the lower-places. Fig. 3 shows a part of the hierarchical structure of place-names in the Japanese postal code number data near 鎮西町.

VII. Evaluation Result

Table II, Fig. 4, and Fig. 5 indicate the result of the evaluation by 4 subjects. For grouping and interpretation of related words (A), groups of related words are obtained in 85% of places. Further, in 37% of places, major parts of related words are interpreted as groups. It can be said that certain feature(s) of those areas are captured in the search result.

On the other hand, for understanding the link between the place and related words (B), seldom links can be found in 36% places. In 83% of places, only one place out of 6 (the original place + 5 lower-places) has a link to the related words.

TABLE II. Result of Evaluation

evaluation	# of (A)	rate of (A)	# of (B)	rate of (B)
≤1.0	12	15%	28	36%
≤2.0	38	49%	37	47%
≤3.0	17	22%	10	13%
≤4.0	9	12%	2	3%
≤5.0	2	3%	1	1%
total	78	100%	78	100%

At present, we will refrain from further analysis of the low scores in evaluation (B). Our subjects, none of whom has detailed knowledge of Karatsu area, may have overlooked the subtle relationships between an upper-place and the related words of the lower-places.

Even when human cannot recognize the link between the lower-place and its related words in blogs, one can find potential related words of the lower-place by our method by taking the upper-place into the account. To confirm appropriateness of our method, the detailed analysis of the blog text will be necessary.

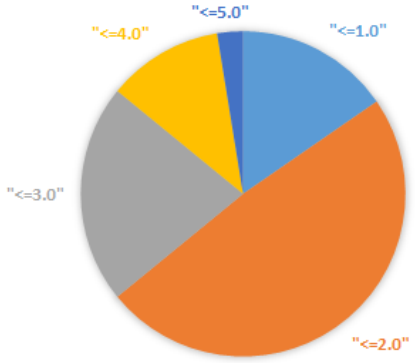

Fig. 4. (A) Evalluation of Related Word

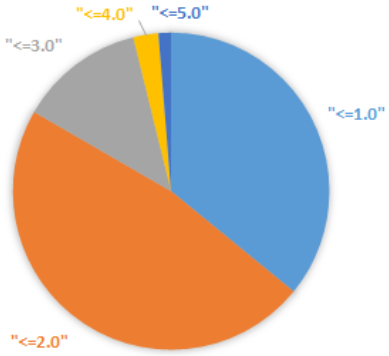

Fig. 5. (B) Evalluation of Place Name

VIII. Conclusion and Further Work

When a visitor comes to the place where he/she doesn't know, "what is the special feature and where" is an important information. To enhance "regional development", it is important to help even a custom business trip visitor to enjoy the special feature of the area.

There are vivid comments based on actual experiences in personal blogs. Those comments are not necessarily on famous places. There may exist unexpected little-known good places near the famous place.

Lively "regional development" can be expected by connecting the special features of the areas currently recognized only as individual isolated points. However, the hierarchical structure of place-names that will be useful in connecting individual points does not appear explicitly in Web information, including blogs etc.

In this paper, we have proposed a technique to find potentially related words which accompany a place using the hierarchical structure in the Japanese postal code system. By expanding place-names using the hierarchical structure of the places, it is possible to guide those who get interested in a specific narrow area to a somewhat wider area.

As an evaluation experiment, we have extracted potential related words of 78 place-names which does not appear

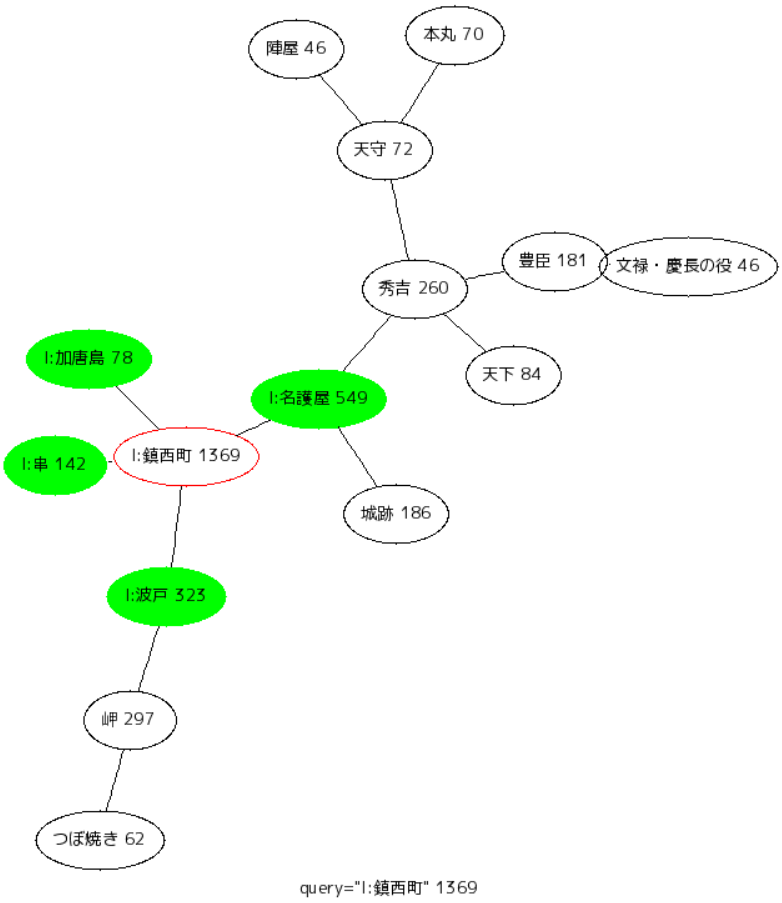

Fig. 2. MindMap of “Chinzei-Chou”

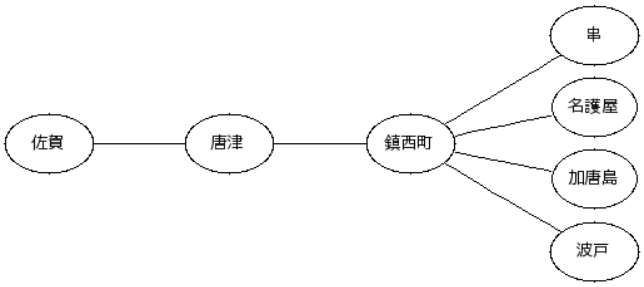

Fig. 3. Hierarchy of Place Name(Chinzei-Chou)

explicitly in blogs. In our experimental data, there are 45,533 blog entries that are relevant to Karatsu and 867 place-names in Saga.

For each 78 name of place, best 20 related words and best 5 related places are extracted and evaluated manually by 4 subjects. The evaluation is based on two viewpoints. That is, (A) whether the related words can be grouped and can be interpreted, and (B) whether the places and related words can be linked. Subjects have evaluated the results in five grades. Related words of 85% places can be grouped and interpreted to a certain degree. In 37% places, major parts of related words can be classified into groups and interpreted. On the other hand, the subjects hardly found links between the places and the related words.

It is necessary to analyze individual blog texts to determine if the links between expanded places and the related words are appropriate or not. The confirmation is a subject of the future work.

The same place-name can be used in several areas to designate the different areas. This is a polysemy problem of place-names that must be taken seriously. There are a number of polysemous place-names in Saga Prefecture analyzed in this paper. Because there are multiple direct upper-places for a polysemous name of a place, it is possible that a feature of

wrong area that has the different upper-place is assigned to the potential related words. Identification of the polysemous place-names places is another subject of the future work.

The method proposed in this paper is useful for tourist boards of local governments and for tour companies. They have strong motivation to inform tourists of special features in the area. From the viewpoint of tourists, it is desirable to know which area is related to the special foods or events. Our method is also applicable for searching of the related places from the keyword about special foods or events.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 24500176.

REFERENCES

[1] J. M. Ruiz-Martinez, D. Castellanos-Nieves, R. Valencia-Garcia, J. T. Fernandez-Brieis, F. Garcia- Sanchez, P. J. Vivancos-Vincente, J. S. Castejon-Garrido, J. B. Camon and R. Martinez-Bejar, Accessing Touristic Knowledge Bases through a Natural Language Interface, Springer LNAI 5465, pp.147-160, 2009

[2] S. Esparcia, V. Sanchez-Anguix, E. Argente, A. Garcia-Fornes and V. Julian, Integrating Information Extraction Agents into a Tourism Recommender System, Proceedings of HAIS2010, Springer LNAI 6077, pp.193-200, 2010

[3] Q. Hao, R. Cai, Ch. Wang, R. Xiao, J.-M. Yang, Y. Pang and L. Zhang, Equip Tourist with Knowledge Mined from Travelogues, Proceedings of WWW2010, pp.401-410, 2010

[4] I. Kinjo and A. Ohuchi, Web data analysis for Hokkaido tourism information, IEICE Technical Report. DE, 101(193), pp.99-104, 2001

[5] H. Nanba, H. Taguma, T. Ozaki, D. Kobayashi, A. Ishino and T. Takezawa, Automatic Compilation of Travel Information from Automatically Identified Travel Blogs, Proc. of the ACL-IJCNLP 2009 Conference, pp.205-208, 2009

[6] T. Nakatoh, C. Yin and S. Hirokawa, Characteristic Grammatical Context of Tourism Information, ICIC Express Letters, 6(3), pp.753-758, 2012

[7] T. Nakatoh and S. Hirokawa, Evaluation of Tourism Resources Extraction based on Japanese Dependency Analysis, Proceedings of AAI2013, pp. 100-103, 2013
¡

[8] C. Fink, J. Piatko, D. Mayfield,D. Chou, T. Martineau, The geolocation of web logs from textual clues, Proc. of 12th IEEE International Conference on Computational Science and Engineering (CSE 2009), Volume 4, No. 5282996, pp.1088–1092, 2009

[9] E. Amitay, N. Har'El, R. Sivan, R., A. Soffer, Web-a-where: Geotagging Web content, Proc. of 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 273–280, 2004

[10] H. Toda, N. Yasuda, Y. Matsuura, R. Kataoka, "Geographic information retrieval to suit immediate surroundings," Proc. of the 17th ACM International Symposium on Advances in Geographic Information Systems, pp.452-455, 2009.

[11] K.A.V. Borges, C.A. Davis Jr., A.H.F. Laender, C.B. Medeiros, Ontology-driven discovery of geospatial evidence in web pages, Springer GeoInformatica 15(4), pp.609–631, 2011.

[12] S. Hirokawa, B. Flanagan, T. Suzuki, C. Yin, Learning Winespeak from Mind Map of Wine Blogs, in S. Yamamoto (Ed.): Proc. HIMI 2014, Part II, LNCS 8522, pp. 383–393, 2014