IJCAT - International Journal of Computing and Technology, Volume 7, Issue 7, July 2020
ISSN (Online) : 2348-6090
www.IJCAT.org
Impact Factor: 0.835

Identification of Geo-Informatics Using Field Associative Words in Blog Search

Abdunabi Ubul

Department of Civil Engineeirng, Shibaura Institute of Technology
3-7-5 Toyosu, Koto-ku, Tokyo 135-8548, Japan

Abstract -This paper summarizes the research results on the method of extracting field associative words and related words from blog search as part of the semantic understanding technique of the blog document. The purpose of this study is to describe the significance of the identification of geoinformatics using field associative words in blog search. In addition, we explain the proposed field associative word method. In this method, we used a blog containing various categories as input to extract information on the ground field, but we can obtain a correct Accuracy rate of 80% or more, and this method is considered to be effective.
.

Keywords: *Blog search, Geo-Informatics, Field association word*

1. Introduction

In recent years, with the spread of the Internet, there have been more opportunities for individual persons to transmit information to the Web. One of the reasons for the increase in the number of opportunities is blog posts. In the blog, the diary is posted as the main content, but there are a lot of frank opinions about them of general person such as the comment on the geoinformatics. These opinions can be used as information and clues about a topic with ground.

On the other hand, if you think about searching a blog containing the necessary information from the reader's point of view, the number of subjects is enormous because of the amount of blogs that exist on the Web and the short update interval. Therefore, when searching for a blog, we generally use the service providing site blog search of the blog. Below, these sites are called blog sites. In the blog site, you can search by keyword for all blogs. In addition, it is also possible to search by following a category that is prepared in the blog site.

However, in the search by keyword, there is a difference in the search results depending on the skill of the user. In addition, it is necessary for the user to search for the blog that the user requests for the search which follows the category. This is because the contents of the blog belonging to the category are not necessarily written in the same sense as the category. Moreover, it is one of the factors that it is written in the content which has no relation with the category at all.

In the research related to the search of the blog, the research on the automatic classification of the blog and the search for related articles is done [1-5]. However, there was no research on the extraction of geoinformatics using field associative words in the search of blogs belonging to a certain category prepared on the blog site. The purpose of this study is to identify geoinformatics using field associative words in the search of blogs belonging to the category. In the search of related blogs, we aim to extract information on accurate ground fields by narrowing down the blogs belonging to the category and searching for related blogs between categories.

2. Outline of Blog Search

When thinking about collecting information about a blog, it is necessary to search the blog on the Web. However, it is not easy to narrow down the target only to the blog. Therefore, using the "search keyword input form" and "blog category" prepared on the blog site, we search for blogs. As a result, instead of searching the blog from the Web, it is possible to search only the necessary blogs from the blog, it is easy to collect information. In the blog site, two kinds of search methods are adopted to search the blog.

2.1 Understanding category search

In order to do the category search, follow the "blog category" that is prepared in the blog site. In the blog site, the article is presented in the blog and "updated order" in "the order in which you are personnel in the category" according to the category. "Person-like order in category" means that the blog site uses the scores it has decided

independently and presents the blog in a high order.

2.2 Keyword Search

In order to perform a keyword search, the user enters the keyword in the "search keyword input form" prepared on the blog site. In the blog site, a full-text search is performed to search for the blog that contains the keyword, and the blog containing the keyword is presented to the user.

3. Field identification using field associative words

One of the commonalities found in research on searching blogs is "identifying the topic of a blog." In this study, we identify topics with ground field when searching for blogs related to the narrowing down of blogs. To identify topics, we use words written on the blog.

3.1 Field associative words

A word that can be associated with a particular field is called a field associative word [6-10]. For example, common nouns such as "subsidence" and " liquifaction " are common words that can be associated with the field of ground.

3.2 Field identification using field associative words

The field identification of the blog using the field associative word is explained by the concrete example. The field-specific flow is as follows.

Step 1. Extraction of field associative words: extract the field associative word from the sentence written in the blog. At this time, we refer to the database of field associative words that have been constructed in advance.

Step 2. Calculating points: A score according to the level is given to the extracted field association word. This is because even in the field associative words that associate the same field, the degree of the field associated with each field associative word is different. The total score for each field is calculated using the score. At this time, the number of appearances of the field association word extracted from the blog is also calculated. Each blog has field association words for various fields. Therefore, it can be said that it is difficult to identify one field from one blog. Therefore, the field in which the blog is written is specified based on the score and the number of appearances given to the field associative word.

Step 3. Target area identification: In field identification, an associative thesaurus is referenced when extracting field associative words. The field is specified by calculating the score of the field associated word that has been extracted. The details of each process in the example of identifying the ground field shown in Table 3.1 are described below.

Table3.1 includes " Ground subsidence ", " Liquifaction", "Soil material ", and " Pillar diagram ". When the total score for the <Ground> is calculated from the number of occurrences and the score of these field associative words, it becomes as follows.

Table 3.1: Results of extraction of field associative words

Field associative words	Field of association	Number of occurrences
Disaster prevention	<Disaster >	1
sand	<Geology>	1
Ground subsidence	<Ground>	1
Liquifaction	<Ground>	1
Soil material	<Ground>	1
Residential ground	<Ground>	1
Pillar diagram	<Ground>	1

When the total score for the < Ground > is calculated from the number of occurrences and the score of these field associative words, it becomes as follows.
Total points :
= ("Ground subsidence": 40 × 1) + ("Liquifaction ": 30 × 1) + ("Groundwater": 30 × 1) + ("Soil material ": 40 × 1) + ("Residential ground": 40 × 1) + ("Boring hole ": 30 × 1) = 200

Similarly, the total number of < Disaster > and < Geology > is calculated. Here, since there are no more than one field associative words for these fields, the total score is 40,30, respectively.

Table 3.2: Total Points Calculated

Field candidates	Total points	Field associative words	Scores	Number of occurrences
<Ground>	200	Ground Subsidence	40	1
		Liquifaction	30	1
		Groundwater	30	1
		Soil material	30	1
		Residential ground	40	1
		Boring hole	30	1
<Disaster >	40	Disaster prevention	40	1
<Geology>	30	Sand	30	1

<Ground>, <Disaster >, and <Geology> are candidates for the field of the blog from the results of calculating the total score shown in Table 3.2. From these candidates, the <Ground> with the highest total score can be identified as the field of this blog. In addition, if there are multiple fields with the highest total score, both fields shall be the field of the blog.

3.3 Field identification in related blog search flow

The search of the related blog is explained by giving the body of the blog as a concrete example. The search flow is as follows.

Step 1. Extraction of non-field associative words: In the search of the blog, the narrowed blog is input. Since the field associative words of the input blog are extracted when the field is identified, the non-field associative words are extracted here.

Step 2. Check with blog database: The 'blog field', 'field associative words and non-field associative words', of the entered blog, is checked against the 'field of blog' of the blog stored in the blog database, 'field associative words and non-field associative words'.

Step 3. Presentation of related blogs: As a result of the verification, the 'field of the blog' matches, and the number of matching 'field associative words and non-field associative words' is obtained from the blog database and presented.

4. Experiments and Considerations

We conducted an experiment to evaluate the method of narrowing down blogs and searching related blogs. In this chapter, we describe the results and consideration of two experiments, the narrowing experiment of the blog and the search experiment of the related blog.

4.1 Experimental Settings

In the experiment, we used 1100 randomly selected blogs [11-13] from 10 categories centered on ground information, geological information, disaster prevention information, etc. Categories are "Ground", "Geology", "Life", "Disaster Prevention", "Earthquake", "Evacuation", "Disaster", "Tsunami", "Weather", and "Nature". The judgment of the blog which was correctly narrowed down was done by person hand, and the following conditions were set as the judgment criterion.

S1: As a result of the field specific of the blog, it becomes the same name as the category, and the content of the blog is written in the same meaning as the category.
S2: As a result of the field specific of the blog, it became a name different from the category, but the content of the blog is written in a different sense from the category.

It is assumed that blogs satisfying the above conditions can be narrowed down correctly. In addition, precision and recall are used as the criteria for evaluating narrowing. The formulas for precision and recall are as follows.

Note that the formulas for precision and recall are shown for each of the above conditions S1 and S2. In the formula for condition S1, the formula for precision S1, recall S1, and conditions S2 are shown. Now, let us call it precision rate S2 and recall rate S2.

$$\text{Precision } S1 = \frac{C}{A} \times 100 \ (\%)$$

$$\text{Recall } S1 = \frac{C}{B} \times 100 \ (\%) \qquad (1)$$

$$\text{Precision } S2 = \frac{F}{D} \times 100 \,(\%)$$

$$\text{Recall } S2 = \frac{F}{E} \times 100 \,(\%) \qquad (2)$$

A: The number of blogs in which the field and category of the identified blog became the same name,
B: the number of blogs that became different in the field and category of the identified blog,
C: the number of blogs narrowed down by the conditions of S1 and S2.
D: Number of blogs with different categories than the field and category of the identified blog
E: Number of blogs that need to be narrowed down by the conditions of S2
F: Number of blogs narrowed down by the conditions of S2

Then, the percentage (Accuracy rate) that narrowed down all the blogs in the category was calculated by the following equation as the correct answer for blogs that met the conditions of S1 or S2.

$$\text{Accuracy} = \frac{\text{conditions of S1 and S2}}{\text{Total number of blogs}} \times 100 \,(\%) \qquad (3)$$

4.2 Experimental Results

The results of the refinement experiment are shown in Table 4.1 to Table 4.3. Table 4.1 is the result of the narrowing down that meets the conditions of S1, and Table 4.2 is the result of the narrowing down that meets the conditions of S2. Table 4.3 is the result of narrowing down all the blogs in the category. The A to F shown in Table 4.1 to Table 4.3 corresponds to the A to F described in the experimental settings in paragraph 4.1.

From the results of Table 4.1 and Table 4.2, blogs belonging to the category are not classified correctly.

Table 4.1: Results of refinements that meet the criteria in (1)

Category	Precision S1 (%)	Recall S1 (%)
Ground	80.4	90.1
Geological	79.3	92.2
Life	72.9	71.5
Disaster	80.8	91.3
Earthquake	84.0	92.4
Evacuation	86.0	92.8
Disaster	79.5	80.0
Tsunami	84.1	90.1
Weather	82.9	90.4
Natural	81.8	90.7
Total	81.2	88.1

Table 4.2: Results of refinements that meet the criteria in (2)

Category	Precision S2(%)	Recall S2 (%)
Ground	82.3	87.6
Geological	79.6	90.8
Life	70.3	72.7
Disaster	83.4	89.9
Earthquake	80.4	93.8
Evacuation	82.6	91.6
Disaster	78.2	83.8
Tsunami	83.8	90.6
Weather	84.6	93.6
Natural	80.6	87.8
Total	80.5	88.2

Table 4.3: Results of narrowing down all blogs in the category

Category	Total number	Number of blogs that fulfill the conditions of S1 and S2	Accuracy rate (%)
Ground	108	88	81.4
Geological	102	82	80.3
Life	134	94	70.1
Disaster	112	89	79.4
Earthquake	113	94	83.1
Evacuation	115	92	80
Disaster	104	90	86.5
Tsunami	106	88	83
Weather	100	83	83
Natural	106	81	76.4
Total	1100	881	80.3

From Table 4.3, we were able to correctly narrow down more than 80% of the blogs in the category, so we were able to confirm that the results were generally good.

4.3 Consideration

We describe the results obtained by the narrowing experiment of the blog. As mentioned above, the recall rate of Table 4.1 is not a good result compared to other fit and recall rates. If you look at the recall rate by category in Table 4.1, you can see that "life" is the lowest compared to other categories. The reason for this is that the field identification by the field association word is not done correctly.

As an improvement measure, it is possible to add a new word when constructing a field associative word database. By using this method, can also be added, and field identification can be performed correctly.

5. Conclusion and future challenges

In this paper, we have explained the identification of geo-informatics identification in blog search using field association words. The blog search in this research is to divide blogs belonging to a category into blogs that are in the same field as the category and blogs that are in a different field, and to search for blogs that are related to each other. In order to realize these, we have explained the method of extracting field association words from a blog and identifying the identification of geo-informatics of the blog. In addition, when searching for related blogs, we explained a method to judge blogs that have similar contents by using the Identification of geoinformatics of the blog, the extracted field associative word, and the non-field associative word that is a noun that cannot associate the field.

In order to confirm the accuracy of these methods, we conducted a blog narrowing experiment and a related blog search experiment. In the blog narrowing-down experiment, we were able to obtain results that were able to correctly narrow down more than 80% of the blogs in the identification of geo-informatics confirmation experiment. In the related blog search experiment, the recall rate of 88% or more was obtained as a whole, and it was confirmed that the related blogs could be searched without omission.

As a future problem about the narrowing of the blog, it is necessary to narrow down the blog using the parts of the blog page rather than using only the sentence written on the blog as a clue. In addition, it is also mentioned that the field identification by prioritizing the field of the field associative word that appears at the beginning of the sentence, and the sentence written on the blog is summarized by the field by considering the position and the distance of the field associative word written in the blog.

As a future problem about the search of related blogs, it is possible to investigate the change in the fit rate by the number of blogs to be stored in the blog database, and to improve accuracy by storing the position and proper noun and ordinary noun information that appeared in the blog of the field associative word and non-field associative words in the blog database. In addition, it is necessary to investigate the change of the fit rate while changing the time of the experiment.

References

[1] Bautin, M., Vijayarenu, L. and Skiena, S.: International Sentiment Analysis forNews and Blogs, Proc. ICWSM, pp.19–26 (2008).

[2] Evans, D.K., Ku, L.-W., Seki, Y., Chen, H.-H. and Kando, N.: Opinion Analysis across Languages: An Overview of and Observations from the NTCIR6 Opinion Analysis Pilot Task, Proc. 3rd Inter. Cross-Language Information Processing Workshop (CLIP2007), pp.456–463 (2007).

[3] Nakasaki, H., Kawaba, M., Utsuro, T. and Fukuhara, T.: Mining CrossLingual/Cross-Cultural Differences in Concerns and Opinions in Blogs, Computer Processing of Oriental Languages: Language Technology for the Knowledge-Based Economy: 22nd International Conference, ICCPOL 2009 (Li, W. and Moll´a-Aliod, D., eds.), Vol.5459, Springer, pp.213–224 (2009).

[4] Wiebe, J., Wilson, T. and Cardie, C.: Annotating Expressions of Opinions andEmotions in Language, Language Resources and Evaluation, Vol.39, No.2-3, pp.165–210 (2005).

[5] Yangarber, R., Best, C., von Etter, P., Fuart, F., Horby, D. and Steinberger,R.: Combining Information about Epidemic Threats from Multiple Sources, Proc.Workshop: Multi-source, Multilingual Information Extraction and Summarization, pp.41–48 (2007).

[6] E.-S. Atlam, G. Elmarhomy, M. Fuketa, K. Morita and J. Aoe, " A New Method For Selecting English Compound Terms and its Knowledge Representation", I Information Processing & Management Journal 2002, Vol.38, No.6, pp. 807-821

[7] Tshering Cigay Dorji, El-sayed Atlam, Susumu Yata, Masao Fuketa, Kazuhiro Morita, Jun-ichi Aoe, "Automatic building of new field association word candidates using search engine", Knowledge and Information Systems, 2011, Vol.27, No.4, pp.141-161.

[8] E.-S. Atlam, G. Elmarhomy, M. Fuketa, K. Morita and J. Aoe, "Automatic building of new field association word candidates using search engine", Information Processing & Management, 2006, Vol.42, No.4, pp.951–962.

[9] A.Ubul, El.Atlam, H. Kitagawa, M. Fuketa, K. Morita, J. Aoe, "An Efficient Method of Summarizing Documents Using Impression Measurements", An Efficient Method of Summarizing Documents Using Impression Measurements, 2013, Vol.32, No.2, pp.371-391.

[10] Abdunabi Ubul, Hidekazu Kakei, Jun-ichi Aoe: Research on Document Summary Generation Using Attribute Information,IJCAT Journal, Vol.1(1), pp.557 - 569 (2014).

[11] Goo Blog, http://blog.goo.ne.jp/.

[12] Ameba Blog, https://ameblo.jp/

[13] Live door Blog, http://blog.livedoor.com/

Author:

Abdunabi Ubul : received his B. Sc. degree in economics and Management information from Xinjiang University, China in 2004. He has received his M. Sc. degree from Department of Economics, Faculty of Integrated Arts and Sciences, University Of Tokushima, Japan in 2008. Received his Ph. D. degree from Department of Information Science and Intelligent Systems. University of Tokushima, Japan in 2012. His research interests include information retrieval, natural language processing and document processing.