The Seventeenth International Symposium on Artificial Life and Robotics 2012 (AROB 17th '12),
B-Con Plaza, Beppu, Oita, Japan, January 19-21, 2012

Proposal of method to extract location-related words and to classify location-dependent information

Masaki Sakata[1], Hiroyuki Nishiyama[1]

[1] Graduate School of Sci.and Tech, Tokyou University of Science
Yamazaki 2641, Noda-shi, CHIBA, 278-8510
(Tel: 81-47122-1106, Fax: 81-47122-1106)
(j7411613@ed.tus.ac.jp)

Abstract: In this study, we collected and analyzed tweets on "Twitter" to classify information of categories related to the location to systematically provide information about the location. This method extracts keywords of a high occurrence ratio (location-related words) appearing in tweets from a set of tweets generated at one location. We hypothesized that tweets including location-related words (location-dependent information) contained information related to the location, and classified the information accordingly. This method enables classifying the information related to the location and providing it to users even from information without an address or location.

Keywords: Twitter, location information, location-related word

1 INTRODUCTION

This century is said to be the era of an advanced information society, in which information technology has been developed, such as computers and networks. With the spread of high-performance mobile devices such as smart phones in recent years, we have become able to transmit and receive information anytime, and anywhere. We are thus able to transmit information easily now but have difficulty extracting and obtaining only the information we want from the vast amount of information available. Therefore, we need the ability to identify information proactively, judge its value, and use it suitably. The high-performance mobile devices mentioned above also have a GPS capability, so we can now transmit and receive location information. However, with this GPS capability, even more information has become available to us, so it is even more important that we selectively acquire only the appropriate information. But we hereby have been more needed to acquire appropriate information from the information became more extensive. In the past, we could acquire appropriate information using a keyword search and link structure. But regarding information related to the location, the intensive and provision has not been systematic currently. However, location information has not been widely provided. Although location information is an important factor in categorizing search results, the supply and demand for information is not sufficiently balanced. In this study, we propose a new method of providing location information. We use "Twitter" as a source that includes location information and classify only the tweets that depend on one location from the number of tweets generated. "Twitter"[1] is used as the information because it is able to provide a number of strings with location information attached, strings are obtained from a broad strata of users, and the freshness of the information is good. In 2010, Arakawa et al. proposed methods of acquiring the standard deviation of latitude and longitude for a tweet group, measuring the locational dependence of keywords from the degree of variation in the tweet group, and performing a two-dimensional depth-first search to extract the area with more than a certain percentage of tweet and extracting some locational dependence for the keyword[2]. That enables quantifying the location dependence of the keyword. In the same year, they conducted research on analyzing the string obtained from the location and time information of tweets, and revealed a correlation between the string and the actual input string[3]. Geocoding is a technique for converting geographical information such as station name and address or place name into latitude and longitude coordinates. Using this technique enables obtaining coordinates from information including place names in the text and linking the textual information and coordinates. As an example, there are services such as "maplog"[4] that enables searching the latest blog posts which are written about the range of the map to display. Also, a method to classify messages by category and aggregate similar messages using spatiotemporal information has been proposed by Yamanaka et al[5].

2 SYSTEM DESIGN

Using the geocoding techniques described in the previous section, the categories of information that can be obtained are limited to geographical information. However, there are other categories of information related to the location such as buildings, traffic, and weather. This study attempts to provide only the information of categories related to the loca-

tion. In this study, the keywords with a high occurrence ratio (location-related words) are extracted from a number of tweets generated on "Twitter" at one location, and location-related words are registered in the database on the server together with location information. We hypothesized that, based on the database created, tweets including location-related words in the text (location-dependent information) contained other information related to the location. Additionally, by repeating the procedure of extracting location-related words and classifying location-dependent information, we thought that extra information (non-location-related words) would be excluded, the number of missed words would be reduced, and the latest information would be provided while responding to location-related words and location-dependent information that varies with time. Also, information is provided by displaying tweets (location-dependent information) in the range of the map for users of this system to display in the browser on the map. The following describes in detail the proposed method for each procedure. In this study, words that have not been determined to be location-related words yet or words for which it is still been unclear whether they are location-related words are referred to as candidate location-related words.

2.1 Extracting location-related words

In this method, tweets including location information are collected. The tweets of Twitter users are collected continuously (can designate a collection area) and are registered in the database together with location information. At the same time, each tweet is morphologically analyzed, and several morphemes generated from a single tweet are registered in the database. Keywords with a high incidence are extracted from a number of tweets generated at one location, and are registered as location-related words in the database. Thus, it is possible to determine location-related words for each area. In addition, location-related words are expected to continue to change constantly with the behavior of Twitter users. However, it is always possible to respond to changing location-related words by constantly repeating the steps from collecting tweets to extracting location-related words, and using only tweets generated in a certain time period. The problem here is how to determine which location tweets are expressed.

2.1.1 Determining the area to which tweets belong

This section describes how to determine which location tweets are indicating. For example, we will consider the situation in Fig. 1. If tweet a was generated at Point X, the tweet could be determined to be indicating Landmark A. A landmark here is a reference point on the map and refers to the name of a station, government office, school, hospital, or post office. In this case, it is possible to identify the area by

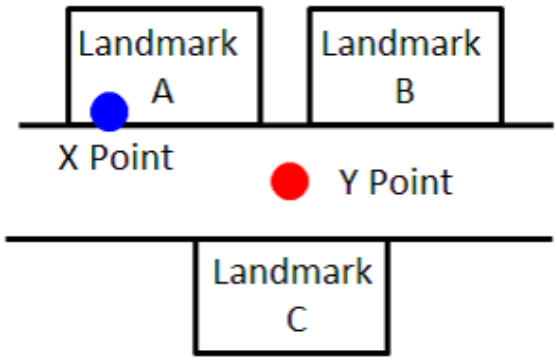

Fig. 1. Subject of tweets: Point X is clear; Point Y is unclear.

performing reverse geocoding. Reverse geocoding is a technique to convert the latitude and longitude coordinates for geographical information such as station name and address or place name. However, if it is unclear which location the tweet is indicating like Point Y, it is necessary to know how to resolve the ambiguity. In this method, the area around the tweet generation spot is defined as the area to which the tweet belongs. If the tweet is generated at the red dot in Fig. 2, the range inside the blue circle around the red dot is the tweet area. If the tweet area is partially determined, it is also possible to define the area which an individual morpheme is extracted from the tweet belongs to is equal to the area which the tweet belongs to. Thus, the tweet area is defined by using reverse geocoding and our proposed method together. One more matter must be defined. The tweet area has been determined, but we must determine how to identify location-related words and the area to which they belong from morphemes extracted from several tweets (candidate location-related words).

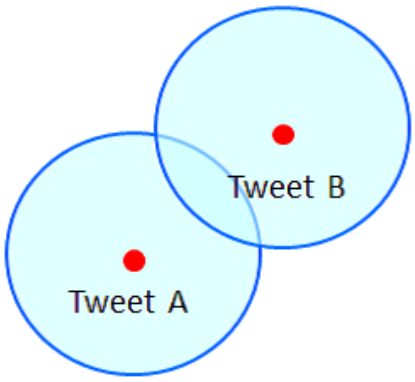

Fig. 2. Area to which tweets belong

2.1.2 Determining location-related words and the area to which they belong

This section describes how to determine the area to which location-related words belong. Even if tweets contain certain

location-related words, they can not be location-independent unless they are tweets generated in the area to which location-related words belong. Therefore, how to determine the area to which location-related words belong is important. We will consider focusing on one keyword. The area to which tweets belong is described in the above section. Here, in an area to which a certain location-related word inherently belongs, the tweets including the location-related word should be somewhat dense. Conversely, in an area to which the location-related words inherently should not belong to, generated tweets including the location-related word are few and sparse. If these sparse tweets are removed, the area to which the location-related word belongs can be specified to some extent. For example, assume that tweets including a certain keyword are generated by many users in the positions depicted in Fig. 3. In the area to which a certain location-related word inherently belongs, some of the areas encompassing individual tweets including the location-related word will overlap. In an area to which the location-related word inherently should not belong, there should be few such overlaps. Therefore, only tweets with a specified threshold number of overlaps with other tweets (tweets in the green area in Fig. 3) are adopted, and the tweets with fewer overlaps are removed. Thus, the area is determined by the density of certain candidate location-related words. Pay attention to still being candidate location-related words at phase to have determined the area. When the area is determined, a set of areas composed of several tweets is found (the set of green circles in Fig. 3). If the number of individual tweets composing this set (Fig. 3; three in this example) exceeds a certain number, the word is determined to be a location-related word. Thus, after determining the area in the candidate location-related words phase, the words are determined to be location-related words depending on the number of tweets composing the area. The next section describes how to classify location-dependent information with these location-related words.

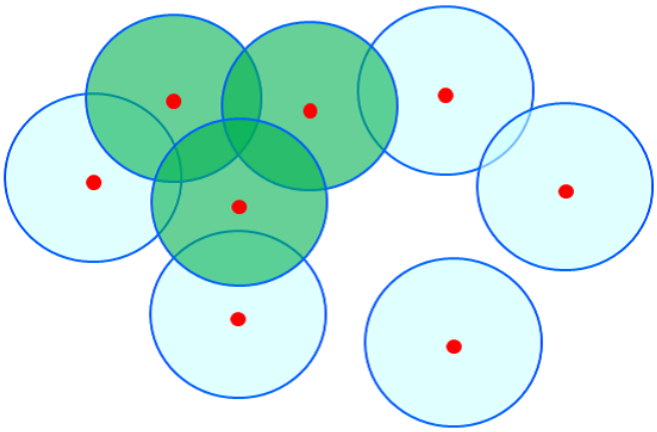

Fig. 3. Area to which location-related words belong

2.2 Classification of location-dependent information

Location-dependent information is classified based on location-related words described in the above section. On the theory that, based on the database created, tweets including keywords determined to be location-related words in the location in the text contained information related to the location, the tweets are acknowledged as location-dependent information. However, this classification rule causes a problem. It can be easy to image that the words expressing a place name and location also become candidate location-related words. In "Twitter," too many tweets express only their own place, such as the Japanese word "nau." Because words expressing a place name and location are also registered as location-related words, tweets expressing only their own place are also acknowledged as location-dependent information. Thus, it is necessary to remove unwanted information from the tweets including location-related words by classifying by means. In this way, location-dependent information is determined from tweets including location-related words in the location in the text by removing unwanted information.

3 IMPLEMENTATION

In this study, Apache is used as a Web server and MySQL is used as the database for the server. The Web server-side programs is implemented using php; the browser-side program, which is a client, is implemented using JavaScript. Tweets on "Twitter" are collected using Twitter API, and location-dependent information is provided to a browser using Google Maps API. Fig. 4 illustrates the system processing flow.

Tweets containing Twitter user location information are collected and registered in the database together with other location information (①). Morphemes generated from the tweets are also registered in the database with location information (②). Location-related words are determined from a set of generated morphemes by the method described in Section 2.1 (③, ④). Using the results, a set of tweets that are already registered in the database is classified (from tweets in the area to which location-related words belong) according to whether it contains location-related words (⑤). Unwanted information described in Section 2.2 is removed from classified tweets by means (⑥). Classification of location-dependent information is complete at this stage, and tweets which include location-dependent information are provided for users of this system by displaying tweets on the map with location information (⑦, ⑧). Fig. 5 presents an example.

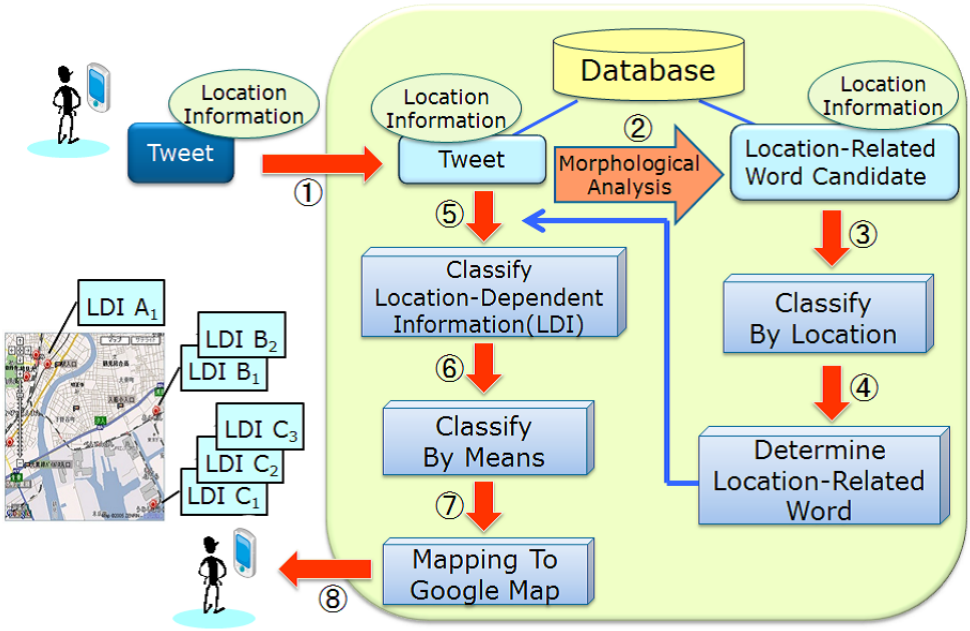

Fig. 4. System flow

This system can be used even from a PC or a smartphone using a browser.

Fig. 5. Example

4 CONCLUSIONS

In this study, we sought to classify only the information of categories related to the location as a systematic method of providing information related to the location. This enables us to classify the information related to the location even from information that cannot be found by conventional methods such as geocoding, not including geographical information such as the address or location. Location-related words change due to changes in the behavior of Twitter users, but if the system is running, this method is able to update location-related words and provide fresh information without the need for analysis. In the future, we will consider extending our method to determine information that users need depending on the content of tweets and location information of Twitter users and to provide location-dependent information focused on the users. We would like to pursue the usefulness of this method further by repeating these extensions of the system.

REFERENCES

[1] Twitter Company:Twitter.http://twitter.com/.

[2] Yutaka Arakawa, Shigeaki Tagashira, Akira Fukuda (2010), Extraction of Location Dependent Words from Twitter Logs (in Japanese). Information Processing Society of Japan SIG Technical Report MBL-55(10):1-6.

[3] Yutaka Arakawa, Shigeaki Tagashira, Akira Fukuda (2010), Relational Analysis between User Context and Input Word on Twitter (in Japanese). Information Processing Society of Japan SIG Technical Report UBI-25(50):1-7.

[4] New Wave Company:maplog.http://maplog.jp/.

[5] Tsutomu Yamanaka, Yuta Tanaka, Yoshinori Hijikata, Shogo Nishida (2010), A Supporting System for Situation Assessment using Text Data with Spatio-temporal Information (in Japanese). journal of Japan Society for Fuzzy Theory and Intelligent Informatics 22(6).