Regional Analysis Based on Location Information and Time Series Change Using Geotagged Tweets

Masaki Endo, Shigeyoshi Ohno
Division of Core Manufacturing
Polytechnic University
Kodaira-shi, Tokyo
e-mail: endou@uitec.ac.jp, ohno@uitec.ac.jp

Masaharu Hirota
Faculty of Informatics
Okayama University of Science
Okayama-shi, Tokyo
e-mail: hirota@mis.ous.ac.jp

Tetsuya Araki, Hiroshi Ishikawa
Graduate School of Systems Design
Tokyo Metropolitan University
Hino-shi, Tokyo
e-mail: araki@tmu.ac.jp, ishikawa-hiroshi@tmu.ac.jp

***Abstract*—Because of the popularization of Social Networking Services (SNSs), it is possible to acquire large amounts of data in real time. For this reason, various studies are being conducted to analyze social media data and to extract real-world events. Among them, a salient advantage of analysis using positional information is that one can accurately extract events from target areas of interest. However, data having position information in social media data remain few: the data amount might be insufficient for analysis. Therefore, we are assessing a method for real-time analysis using data with location information that has been accumulated over a certain period of time. In this research, we are studying a method of regional analysis by position information and time series change using tweets with Twitter position information. Herein, we explain the results obtained from area analysis using the proposed method.**

Keywords-location information; time series; Twitter.

I. INTRODUCTION

In our everyday life, because of the wide dissemination and rapid performance improvement of various devices such as smartphones and tablets, diverse and vast data are generated on the web. SNSs have become especially popular because users can post data and various messages easily. Twitter [1], an SNS that provides a micro-blogging service, is used as a real-time communication tool. Numerous tweets have been posted daily by vast numbers of users. Twitter is therefore a useful medium to obtain, from a large amount of information posted by many users, real-time information corresponding to the real world.

By analyzing the information sent by these SNSs, the possibility exists of obtaining useful information in real time. We are conducting research related to providing tourist information to travelers. Therefore, this study specifically examines the provision of real-time sightseeing information.

Herein, we describe the provision of information to tourists using web contents. Such information is useful for tourists, but providing timely and topical travel information entails high costs for information providers because they must update the information continually. Today, providing reliable information related to local travel is not only strongly demanded by tourists, but also by local governments, tourism organizations, and travel companies, which bear high costs of providing such information.

For that reason, providing current, useful, real-world information for travelers by ascertaining changes of information according to seasons and time zones of the tourism region is important for the travel industry. It is possible to disseminate information using the popular SNS, but organizations that can actually do the work are limited by human resources and cost. Therefore, analysis using an SNS that can provide useful data leading to real-time information provision is one means of overcoming this difficulty.

To solve this problem, much research to analyze SNS data is currently being conducted. Research using Twitter is one branch of investigation. Because tweets comprise short sentences, a location can be estimated if a tweet includes the place name and the facility name, but if such information is not included, identifying the location from a tweet might be difficult. For this reason, research using tweets with location information or tweets which give location information in the tweet itself is being conducted. Because geotagged tweets can identify places, they are effective for analysis. Nevertheless, few geo-tagged tweets exist among the total information content of tweets. It is therefore not possible to analyze all regions. For that reason, we also use geotagged tweets to conduct research using information interpolation to estimate the position around the area that is not specified by the position information [2].

Currently, we are considering a method for real-time analysis by collecting temporal and spatial information for a certain period of time using only geotagged tweets, which are said to have a small amount of information. This report presents an experimental approach.

 ISBN: 978-1-61208-697-2

The remainder of the paper is organized as follows. Section II presents earlier research related to this topic. In Section III, we propose a method for real-time analysis using data collected for a certain period. Section IV describes experimentally obtained results for our proposed method and a discussion of the results. Section V presents a summary of the contributions and expectations for future work.

II. RELATED WORK

Various studies are being conducted using SNS position information. Omori et al. [3] proposed a method to extract geographical features such as coastlines using tags of photo sharing sites with geotags. Sakaki et al. [4] proposed a method to detect events, such as earthquakes and typhoons based on a study estimating real-time events from Twitter. By analyzing the Twitter text stream, Pratap et al. [5] proposed a solution to optimize traffic control by considering previous traffic analysis methodology and social data in real time. Various analytical methods have been proposed for analyzing SNS using position information and time series information. However, analysis of data in which large amounts of position information and time series information exist is mainly addressed. Few research efforts examine information using only a few data.

Some research has examined visualization. Nakaji et al. [6] proposed using a geotagged and visual feature of a photograph and suggested a way to select photographs related to a given real event from geotagged tweets. They developed a system that can visualize real-world events on online maps. In the GeoNLP Project [7], we are developing a geotagging system that extracts location descriptions such as place names and addresses contained in natural language sentences. The system provides metadata about where the sentences are descriptions. It is also offered as open source software. These studies are very useful for extraction of specific designated events and for analysis of preregistered places. However, another discussion must be held about automatically extracting events and identifying new places.

As described above, conducting research using geotagged tweets for places with small information amounts and new events and places represents a new approach. Therefore, this research was conducted to identify events and places in real time using accumulation of information and differences in space-time space.

III. OUR PROPOSED METHOD

This section presents a description of a method for target data collection using our method of real-time analysis with position information and time series information.

A. Data collection

Here, we explain the data collection target for this research. Geotagged tweets sent from Twitter are the collection target. The range of geotagged tweets includes the Japanese archipelago (120.0° E ≤ longitude ≤ 154.0° E and 20.0° N ≤ latitude ≤ 47.0° N) as the collection target. Collection of these data was done using a streaming Application Programming Interface (API) [8] provided by Twitter Inc.

Next, we describe the number of collected data. According to a report by Hashimoto et al. [9], among all tweets originating in Japan, only about 0.18% are geotagged tweets: they are rare among all data. However, the collected geotagged tweets number about 70,000, even on weekdays. On some weekend days, more than 100,000 such messages are posted. We use about 423 million geotagged tweets from 2015/2/17 through 2018/12/26. Therefore, we examined 19 million geotagged tweets in Tokyo for these analyses.

B. Preprocessing

This section presents preprocessing after data collection. Preprocessing includes reverse geocoding and morphological analysis, with database storage for data collected using the process.

Reverse geocoding identified prefectures and municipalities by town name using latitude and longitude information from the individually collected tweets. We use a simple reverse geocoding service [10] available from the National Agriculture and Food Research Organization in this process: e.g., (latitude, longitude) = (35.7384446N, 139.460910W) by reverse geocoding becomes (Tokyo, Kodaira-Shi, Ogawanishi-machi 2-chome). In addition, based on latitude and longitude information of the collected tweets, data are accumulated by the same place. As data accumulate, the data are saved in mesh form as time elapses.

Morphological analysis divides the collected geo-tagged tweet morphemes. We use the “Mecab” morphological analyzer [11]. As an example, “桜は美しいです” (“Cherry blossoms are beautiful.” in English) is divisible into "(桜 / noun), (は / particle), (美しい / adjective), (です / auxiliary verb), (。 / symbol)".

Preprocessing performs the necessary data storage from the result of data collection, reverse geocoding, and morphological analysis processing. Data used for this study are the tweet ID, tweet posting time, tweet text, morpheme analysis result, latitude, and longitude.

C. Analysis method

This section presents a description of the method of real-time analysis using position information and time series information.

The analytical method we proposed has the following three stages.

1. Extraction of places by fixed point observation
2. Analysis considering the time series based on 1
3. Analysis using co-occurring words of 2

Therein, 1 is an estimate of the location derived from stationary observation. At such spots, even in places with few tweets, one can discover the location through long-term observation. This method does spot extraction by adding geotagged tweets including specific keywords for long periods at every latitude and longitude.

As presented above, 2 is a method of extracting new spots using spot information accumulated over a long period

as a baseline, by consideration of the time series and finding differences.

As shown above, 3, time series analysis including co-occurrence words in 2 and for keywords used in 1 is performed using the results of morphological analysis of tweets. It is a method used because differences in latitude, longitude, and time series alone might be insufficient to extract differences in data.

Through analyses using these proposed methods, we aim to capture real-time changes in specific areas.

IV. EXPERIMENTS

This section presents a description of a real-time analysis experiment using the method proposed in Section III.

A. Dataset

Datasets used for this experiment were collected using streaming API, as described for data collection in Section III-A. Data are geo-tagged tweets from Tokyo during 2015/2/17 – 2018/12/26. The data include about 19 million items. We use these datasets for experiments to conduct the three methods proposed in Section III.

B. Experimental method

In this section, experiments using the proposed method shown in Section III are described in 1 to 3.

1. Extraction of places by fixed point observation

   This experiment was conducted for Takao-machi, Hachioji, Tokyo: an area of about 4 km east–west and about 2.5 km north–south, as shown in Figure 1. Experimentally obtained results described later are included within the thick frame depicted in Figure 1. For this area, we conducted an extraction experiment with the target word as "cherry blossom" in Japanese as "桜", "さくら", or "サクラ". In all, 65 tweets were found to include a target word.

2. Analysis in considering of time series based on 1

   This experiment was conducted for 4-chome, Myojin-cho, Hachioji city, Tokyo: an area of about 700 m east–west, and about 600 m north–south, as shown in Figure 2. Experimentally obtained results described later are included within the thick frame in Figure 2. For this area, we conduct an extraction experiment with the target word as "ramen" in Japanese as "ラーメン", "らーめん", or "拉麺". In all, 301 tweets were found to include a target word.

3. Analysis using co-occurring words of 2

   This experiment was conducted for Marunouchi 1-chome, Chiyoda-ku, Tokyo: an area of about 1 km east–west and about 1 km north–south, as shown in Figure 3. Experimentally obtained results described later are included within the thick frame in Figure 3. For this area, we set the target word as "ramen", as in the second experiment, in Japanese as "ラーメン", "らーめん", or "拉麺". In all, 6,979 tweets included a target word. As words co-

Figure 1. Target of Takao-machi, Hachioji City.

Figure 2. Target of Myojin-cho 4-chome, Hachioji City.

Figure 3. Target area of Marunouchi 1-chome, Chiyoda-ku.

occurring after the object, we target tweets including "𠮷蔵" and "玉", which represents the ramen shop name; the respective numbers of tweets were 273 and 31.

C. Experimental result

In this section, the results of 1–3 experiments explained in the previous section are presented.

1. Extraction of places by fixed point observation

The distributions of geotagged tweets in Takao-machi, Hachioji City including cherry blossoms obtained in the experiment are shown in Figure 4 in 2017 and in Figure 5 in 2018. The interior are of the bold frame in Figure 1 is shown in the table. It is about 265 m measured east–west and about 85 m measured north–south. The closer the color of the cell is to black, the more data are shown.

Data extracted for this experiment were very few: 65 for the entire collection period. However, in 2017 and 2018, we confirmed tweets to JR Takao Station, Takao Yamaguchi Station, Takao Station of Ropeway, and Takao Mountain. The correlation coefficient between the extracted spots in 2017 and 2018 was 0.769: high positive correlation was found. Extraction of spots is possible even with few data. Moreover, various spots can be extracted when using longer periods. Therefore, fixed point extraction of sightseeing spots is regarded as possible through continuing observation of geotagged tweets.

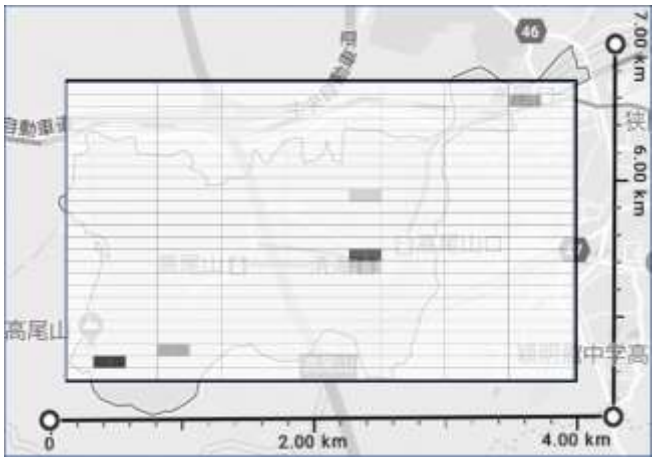

Figure 4. Number of Tweets including Target Words in Takao-machi in 2017.

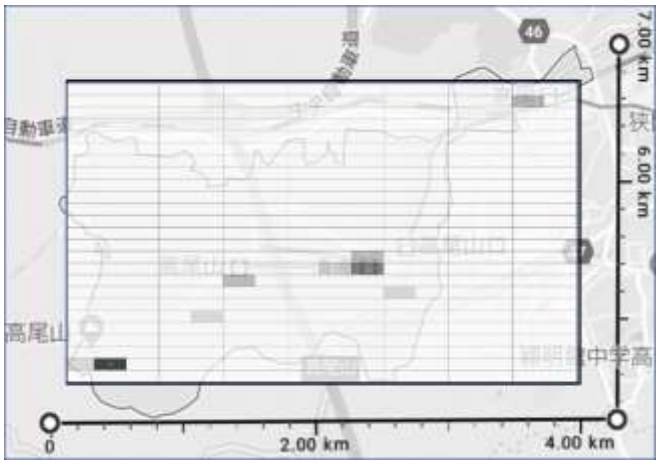

Figure 5. Number of Tweets including Target Words in Takao-machi in 2018.

2. Analysis in considering of time series based on 1

The results for distribution of the geotagged tweets of Myojin-cho 4-chome, Hachioji City including the ramen obtained in the experiment are shown in Figure 6 in 2016, Figure 7 in 2017, and Figure 8 in 2018. The area inside of the bold frame in Figure 2 is shown in the table. It is about 102 m measured east–west and about 39 m north–south. The closer the color of the cell is to black, the more data are shown.

In this experiment, 301 data were extracted in all collection periods. The target area has three ramen shops, which can be extracted from the table of each year. Furthermore, in 2018, a new ramen shop opened at one point; in 2018 we extracted a new spot (latitude, longitude) = (35.69540009399, 139.345001221). For this reason, the correlation coefficient between data of 2016 and 2017 was 0.991, a high positive correlation was obtained. Nevertheless, the correlation coefficient between 2017 and 2018 was only a weak positive correlation of 0.161. These results demonstrated the possibility of extracting new spots by fixed point observation.

3. Analysis using co-occurring words of 2

Distributions of geotagged tweets of Chiyoda-ku including Ramen obtained in the experiment are shown in Figure 9 for 2017 and Figure 10 for 2018. The interior area of the bold frame in Figure 3 is shown in the table. It is about 80 m measured east–west and about 25 m north–south. The closer the color of the cell is to black, the more data are shown.

The data extracted in this experiment were 6,979 in all collection periods. The area of about 540 m east–west and about 540 m north–south in the frame of the heavy line in Figure 3 is a spot called Tokyo Ramen Street, with eight ramen shops. Therefore, as an analytical example using words co-occurring in the target word, Figure 9 and Figure 10 are experimentally obtained results including A="𠮷蔵" co-occurring in the target word and B = "玉".

The ramen shop including A opened in September 2013 and closed in September 2018. The ramen shop including B opened on October 30, 2018.

Therefore, although a difference exists in the number of data, it can be extracted as a spot. However, the closed A information is unsuitable for use as real-time information. The results of analysis considering the time series information are shown in Figure 11. From these results, tweets containing A are not extracted after 2018/9. However, tweets containing B have been extracted since 2018/10. Therefore, by considering the time series in addition

 ISBN: 978-1-61208-697-2

to latitude and longitude information, one can omit the old information in addition to extracting new spots. This result confirmed the possibility of realizing real-time information extraction.

Figure 6. Number of Tweets including Target Words in Myojin-cho in 2016.

Figure 7. Number of Tweets including Target Words in Myojin-cho in 2017.

Figure 8. Number of Tweets including Target Words in Myojin-cho in 2018.

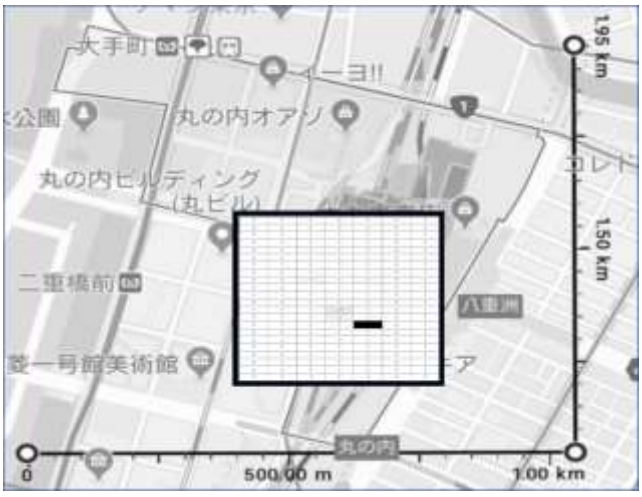

Figure 9. Number of Tweets including “A” Co-occurring in Target Words in Marunouchi 1-chome.

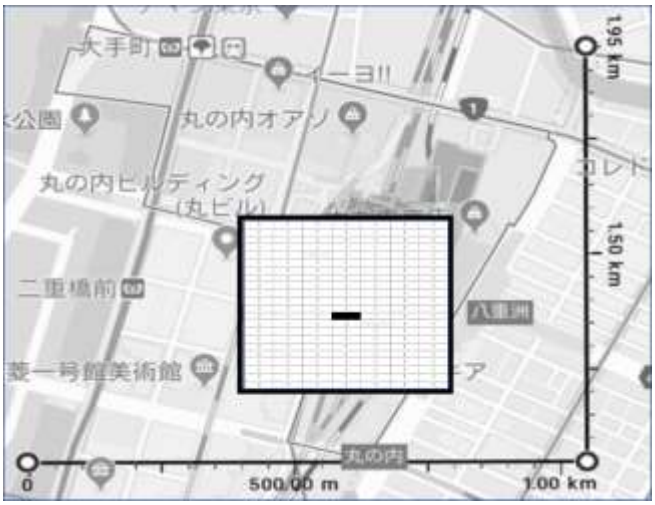

Figure 10. Number of Tweets including “B” Co-occurring in Target Words in Marunouchi 1-chome.

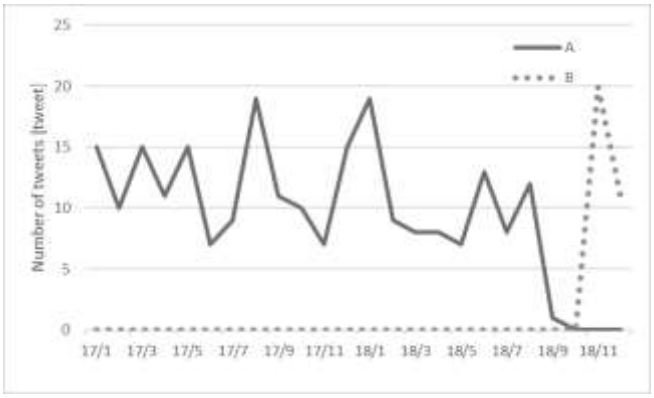

Figure 11. Trends in number of tweets including target words.

V. CONCLUSION

As described in this paper, we evaluated a regional analysis method based on positional information and time series change using tweets with Twitter location information to provide real-time information.

To conduct real-time regional analysis, after proposing a method using geotagged tweets' position information and time series information, we showed experimentally obtained results obtained using that method. Experiment results demonstrated that, even when geotagged tweets were few, spots could be extracted using position information with long-term accumulation. We also confirmed that new spots can be extracted by conducting time series analysis of spot information of position information. Furthermore, using morphological analysis results of tweets, we demonstrated the possibility of analyzing spots, even in densely populated areas with a large amount of information.

Results show that we demonstrated the usefulness of SNS for providing real-time information. Future studies will examine other methods using machine learning to establish even more effective methods.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grants No. 18K13254, 16K00157, and 16K16158, and a Tokyo Metropolitan University Grant-in-Aid for Research on Priority Areas “Research on social big data.”

REFERENCES

[1] Twitter. *It's what's happening.* [Online]. Available from: https://Twitter.com/ 2015.02.15

[2] M. Endo, S. Ohno, M. Hirota, D. Kato, and H. Ishikawa, “Examination of Best-time Estimation for Each Tourist Spots by Interlinking using Geotagged Tweets,” International Journal on Advanced in Systems and Measurements (IARIA), vol. 10, No 3&4, pp. 163-173, 2018.

[3] M. Omori, M. Hirota, H. Ishikawa, and S. Yokoyama, “Can geo-tags on flickr draw coastlines?,” In Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '14). ACM, pp. 425-428, 2014.

[4] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes Twitter users: real-time event detection by social sensors,” WW W 2010, pp. 851–860, 2010.

[5] A. R. Pratap, J. V. D. Prasad, K. P. Kumar, and S. Babu, "An investigation on optimizing traffic flow based on Twitter Data Analysis," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), pp. 320-325, 2018.

[6] Y. Nakaji and K. Yanai, “Visualization of Real-World Events with Geotagged Tweet Photos,” 2012 IEEE International Conference on Multimedia and Expo Workshops, pp. 272-277, 2012.

[7] GeoNLP Project. *A place name information processing system which maps sentences automatically.* [Online] Available from: https://geonlp.ex.nii.ac.jp/ 2019.02.14

[8] Twitter Developers. *Twitter Developer official site*. [Online] Available from: https://dev.twitter.com/ 2015.02.15

[9] Y. Hashimoto and M. Oka, “Statistics of Geo-Tagged Tweets in Urban Areas (<Special Issue>Synthesis and Analysis of Massive Data Flow),” JSAI, vol. 27, No. 4, pp. 424–431, 2012 (in Japanese).

[10] National Agriculture and Food Research Organization. *Simple reverse geocoding service*. [Online]. Available from: http://www.finds.jp/wsdocs/rgeocode/index.html.ja 2015.04.10

[11] MeCab. *Yet Another Part-of-Speech and Morphological Analyzer*. [Online]. Available from: http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html 2015.04.15

 ISBN: 978-1-61208-697-2