Regular Paper

Analytical Method using Geotagged Tweets Developed for Tourist Spot Extraction and Real-time Analysis

Masaki Endo[*], Munenori Takahashi[**], Masaharu Hirota[***],
Makoto Imamura[****], and Hiroshi Ishikawa[*****]

[*] Division of Core Manufacturing, Polytechnic University, Tokyo, Japan
[**] Graduate Student in Electro-info, Polytechnic University, Tokyo, Japan
[***] Faculty of Informatics, Okayama University of Science, Okayama, Japan
[****] School of Information and Telecommunication Engineering, Tokai University, Tokyo, Japan
[*****] Graduate School of System Design, Tokyo Metropolitan University, Tokyo, Japan
endou@uitec.ac.jp

Abstract – The popularization of social network services (SNSs) has made it 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 information about 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. We propose real-time analysis of tourist information using a simple method that incorporates a moving average of geotagged tweet messages on Twitter (Twitter Inc.), a popular SNS. Nevertheless, social media data that include position information are few: the data might be insufficient for analysis. Therefore, we are assessing a real-time analytical method using appropriate data accumulated over a longer period of time. For this study, we detect tourist spots using a regional analysis method that uses location information from Twitter tweets and time-series changes. The experimentally obtained results demonstrate that our method is useful as a complement to our previously proposed moving-average method. As described herein, we propose a method of extracting tourist attractions from the accumulated tweets and report the results of analysis of the extracted destinations. Herein, we explain results of sightseeing spot analysis of cherry blossoms obtained using geotagged tweets from Tokyo.

Keywords: big data, location information, social media, time series, Twitter

1 INTRODUCTION

From everyday life, because of the wide dissemination and rapid performance improvement of various devices such as smartphones and tablets, vast amounts of diverse data are generated and transmitted via the internet. Social network services (SNSs) have become especially popular media because users can post data and various messages easily. Twitter (Twitter Inc.)[1], a popular SNS that provides a micro-blogging service, is used as a real-time communication tool. Numerous tweets are posted daily by vast numbers of users worldwide. Twitter is therefore a useful medium to obtain, from huge amounts of information posted by many users, real-time information corresponding to the real world.

By analyzing the information distributed by these SNSs, useful information is obtainable in real time. We are conducting research related to provision of 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. However, providing timely and topical travel information entails high costs for information providers because they must update the information continually. Today, reliable information for local travel is demanded strongly not only by tourists, but also by local governments, tourism organizations, and travel companies, which are all burdened by the high costs of providing such information.

For the reasons explained above, ascertaining changes of information according to seasons and time zones of the tourism region and then providing current, useful, real-world information for travelers is important for the travel industry. Disseminating information using popular SNSs can be done, but organizations able to do such work are constrained by human resource and cost limitations. Analysis using an SNS can overcome these difficulties by providing useful data facilitating real-time information provision.

To accomplish this task, much research is currently underway to analyze SNS data. Research using Twitter is one branch of investigation. Tweet messages are short sentences from which a location can be inferred if a tweet includes a place name or a facility name. If such information is not included in the message, then identifying a location from a tweet might be difficult. Therefore, research is underway to use tweets with location information or tweets which give location information in the tweet itself. Geotagged tweets can identify places: they are useful for analysis. Nevertheless, few geo-tagged tweets exist among the total information contents of tweets. Therefore, analyzing all regions is not possible. We also use geotagged tweets to conduct research using information interpolation to infer positions around an area that are not specified by position information [2].

Currently, the authors are assessing a real-time analytical method that requires collection of temporal and spatial information from geotagged tweets over a period of time. This report presents this experimental approach to use small

amounts of information accumulated over longer periods. However, earlier methods are based on the assumption that the traveler knows the tourist spot because the analysis is based on a destination-specific analysis. Although this is effective in the case of familiar tourist spots, it is difficult to use this method in areas that are being visited for the first time or in places for which no prior knowledge exists. As described in this paper, we propose a method for extracting tourist attractions and for analyzing them in real time in areas with tourist attractions for which no prior information is known to the user. The proposed method is one way to help tourists to collect information related to sightseeing in areas for people with no prior information related to the sightseeing spots.

The remainder of the paper is organized as explained hereinafter. Chapter 2 presents earlier research efforts exploring this topic. Chapter 3 explains a real-time analytical method using data collected for a certain period. Chapter 4 describes experimentally obtained results for our proposed method, with related discussion. Chapter 5 presents a summary of the study contributions and indicates some expected avenues of future work.

2 RELATED WORK

Various studies are being conducted using SNS position information. Omori et al. [3] described a method of extracting geographical features such as coastlines using tags of photo-sharing sites with geotags. Sakaki et al. [4] assessed 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] explained a solution to optimize traffic control by incorporating earlier traffic analysis methodologies and complementary real-time social data in one analytical system. Various analytical methods have been proposed for analyzing SNS data using position information and time series information. However, those studies mainly address analysis of data for which large amounts of position information and time series information exist. Few research efforts have examined information using only a few data.

Some research efforts have examined visualization. Nakaji et al. [6] proposed the use of a geotagged and visual features of photographs. They suggested a way to select photographs related to a given real event from geotagged tweets. Their developed system can visualize real-world events on online maps. Through the GeoNLP Project [7], we are developing a geotagging system that extracts location descriptions such as place names and addresses included in natural language sentences. Offered as open source software, it provides metadata of places described in text. Although these studies are very useful for extraction of specific designated events and for analysis of preregistered places, further discussion must be held about automatic extraction of events and identification of new places.

In light of the efforts described above, the present study using geotagged tweets for places with small information amounts and new events and places represents a new approach. This research was conducted to achieve real-time identification of events and places in space–time space based on accumulated information and differences.

3 PROPOSED METHOD

This chapter presents a description of real-time analysis of position information and time series information as a target data collection method.

3.1 Data Collection

Here, we explain the data collection target for this research. Geotagged tweets distributed through 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. These data were collected 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, constituting about 423 million geotagged tweets from 2/17/2015 through 12/26/2018. For these analyses, we examined 19 million geotagged tweets from Tokyo.

3.2 Preprocessing

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

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

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), and (。 / symbol)".

The preprocessing involves storing the necessary data based on the results of data collection, reverse geocoding, and morphological analysis process. Data used for this study are the tweet ID, tweet posting time, tweet text, morpheme analysis result, latitude, and longitude.

3.3 Analytical Method

This chapter 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 two stages.

Stage 1. Extraction of places by fixed point observation

The steps in Stage 1 are the following.

1. After a user specifies the two points of latitude and longitude in the southeasternmost and northwestern regions of the rectangle to collect tourist attraction information, keywords to be extracted (e.g. cherry blossoms) are input.

2. The specified area is divided into 25 and 16 equal parts east–west and north–south, respectively.

3. Tweets in the mesh including keywords are extracted. Then the numbers of tweets for meshes are observed.

Stage 2. Analysis considering the time series based on Stage 1

The steps in Stage 2 are explained below.

1. Numbers of tweets observed per mesh in Stage 1 are analyzed by year and month. At this stage, we manually designate them according to the tourism keyword event. For example, for cherry blossoms, the analysis will be performed on a monthly basis.

2. An event will be determined as a seasonally appropriate tourist destination if it appears only at a specific time of year. Then, using our existing method [12], we make the spot a target for real-time analysis to ascertain whether it is currently in its best season.

Therein, Stage 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 accomplishes spot extraction by accumulating geotagged tweets including specific keywords over long periods at every latitude and longitude. Concretely, one selects an arbitrary area to be observed on the map. Then the selected area is divided equally into north–south and east–west meshes. Although another method of meshing using JISX0410 standardized regional mesh codes exists, an arbitrary area was selected and meshed by hand as an experiment. By long-term accumulation of numerous tweets within this meshed area over a long period, one can infer that the place is a tourist spot. Stage 2 is extraction of new spots using spot information accumulated during a long period as a baseline, by considering the time series, and by finding differences. Through analyses using these proposed methods, we aim to capture real-time changes in specific areas. This goal was achieved by checking the change of specific keywords in the chronological order of each year for the selection in Stage 1. With Stage 2, use of only those tweets with location information was sufficient to extract the spots automatically. The extracted spots can be judged in real time using the estimation method we have developed to date. Through analysis using these proposed methods, we aim to ascertain real-time changes in tourist spots in a specific region.

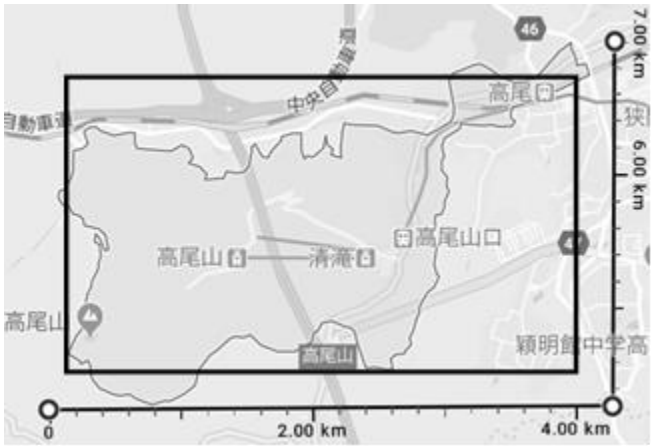

Figure 1: Target of Takao-machi, Hachioji City.

4 EXPERIMENTS

This chapter presents a description of a real-time analysis experiment that was conducted using the method explained in Chapter 3.

4.1 Dataset

Datasets used for this experiment were collected using streaming API, as described for data collection in Sec. 3.1. The data are geo-tagged tweets from Tokyo during 2/17/2015 – 7/29/2019, including about 48 million items. We use these data for experiments to assess the two methods.

4.2 Experiment Method

Experiments designed to assess the proposed method described in Chapter 3 are explained in Sec. 4.2.1 – Sec. 4.2.3.

4.2.1 Extraction of Places by Fixed Point Observation on Stage 1

We conducted a preliminary experiment to ascertain whether spots can be found from the collected tweets, or not. 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 presented in Fig. 1. Experimentally obtained results described later are included within the thick frame depicted in Fig. 1. As a process for Stage 1, we used a mesh divided into 25 equal north–south mesh sections and 16 equal east–west mesh sections. 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.

4.2.2 Time Series Analysis of Stage 2

Spot extraction was performed using time series analysis of Stage 2 for the area shown in Sec. 4.2.1. This analysis is aimed at adjusting the units of the period in the future automatically according to event characteristics, to reflect monthly, weekly, and hourly scheduling. Cherry blossoms bloom once each year. Therefore, we applied a time-series analysis for each year and analyzed them by year. In addition, the analysis for Takao-machi, Hachioji City, Tokyo was conducted for each mesh section in Sec. 4.2.1.

4.2.3 Experiment in Tokyo

We conducted an experiment examining tweets from Tokyo to confirm the accuracy of the proposed method presented in Sec. 4.2.1 and Sec. 4.2.2. This experiment was conducted to

check the accuracy of the spots extracted by Sec. 4.2.1 and Sec. 4.2.2. For comparison, this time, the correct answer spots were 64 spots corresponding to cherry blossom viewing spots in Tokyo, which were listed in the Walker + 2019 edition of Cherry-blossom viewing in Japan. Then, we checked whether known tourist spots can be extracted using the proposed method. The following existing methods [12] were used to find the correct spots.

As a determination method, a 3D map was generated using a spreadsheet program (Excel 2016; Microsoft Corp.). A heat map was used to display the results. A square 100-m on a side was prepared as a spot point centered on the latitude and longitude reported by Walker +. A 1 km diameter circle was also prepared. Next, the following three spot determination methods were used.

1. Places indicated by many tweets are included in the square.
2. Places indicated by many tweets are included in the circle.
3. Places indicated by medium numbers of tweets are in the circle.

Judgment was made based on these three judgment criteria.

4.3 Experiment Results

This chapter presents results obtained from the experiments described in Sec. 4.2.1 – Sec. 4.2.3.

4.3.1 Experiment in Takao-machi, Hachioji, Tokyo

The distribution of geotagged tweets from Takao-machi, Hachioji, including cherry blossoms, as obtained from the experiments described in Sec. 4.2.1 and Sec. 4.2.2, are presented in Fig. 2 for 2017 and Fig. 3 for 2018. The interior area of the bold frame in Fig. 1 is described in the table: it is about 265 m measured east–west and about 85 m measured north–south. This area is obtained by dividing the maximum value and the minimum value of latitude and longitude into 25 and 16, portions respectively, for geotagging tweets. For the current experiment, the number of divisions was specified manually because the target area differs for each experiment. The more closely the color of the mesh section approaches black, the more data are associated with that area.

The tweet data extracted for this experiment were very few: 65 for the entire collection period. However, in 2017 and 2018, we confirmed tweets at JR Takao Station, Takao Yamaguchi Station, Takao Ropeway Station, and Takao Mountain. The correlation coefficient between the extracted spots in 2017 and 2018 was 0.769: high positive correlation was found. The correlation coefficient is calculated using equation (1).

$$\mathrm{Correl}(X-Y)=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}} \quad (1)$$

$$\mathrm{X}: Number\ of\ Tweets\ of\ 2017_{(latitude,longitude)}$$

$$\mathrm{Y}: Number\ of\ Tweets\ of\ 2018_{(latitude,longitude)}$$

Results of the time-series analysis by year for the data collection period in the experimental area are portrayed in Fig. 4, which shows that most of the tweets were collected in March and April. Therefore, one can identify the cherry blossom viewing related spots and analyze the cherry blossom season. Although some tweets are related to cherry blossoms during the season of autumn leaves, the tweets generally include messages about "cherry blossoms of the four seasons" and messages about seasons when people view cherry blossoms. Although accuracy improvement through noise removal is necessary, we limited the recommended seasons even for small numbers of tweets.

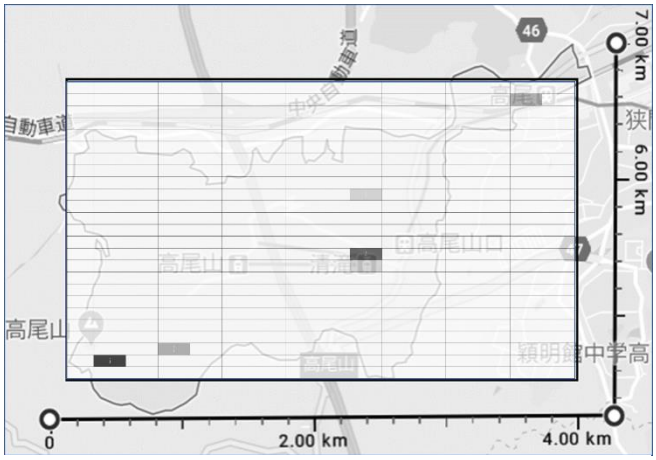

Figure 2: Number of Tweets including target words in Takao-machi in 2017.

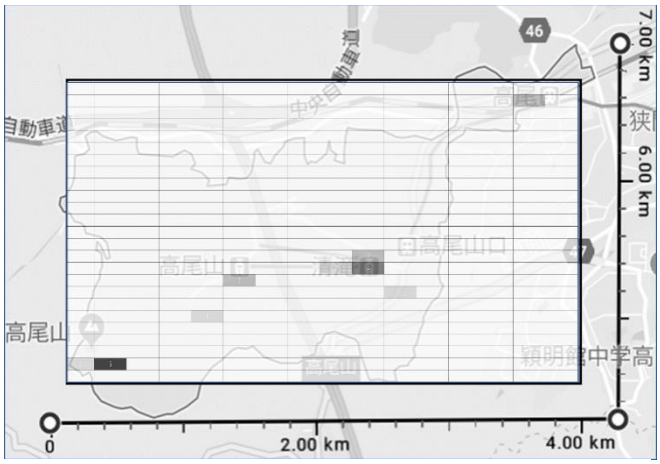

Figure 3: Number of Tweets including target words in Takao-machi in 2018.

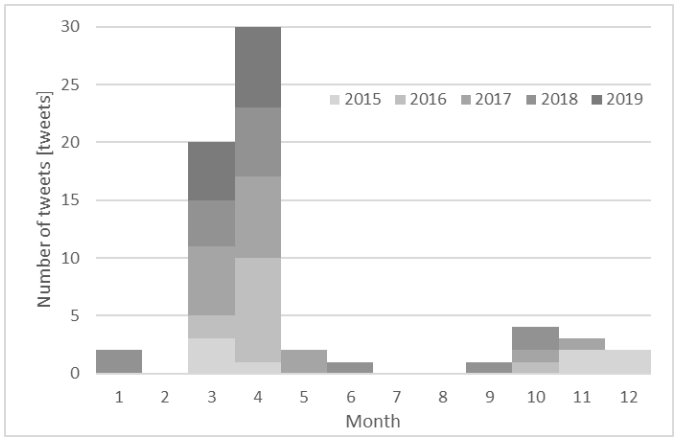

Figure 4: Results of the time-series analysis in Takao-machi.

Spot extraction 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.

4.3.2 Spot Extraction Experiment Results in Tokyo

We present results obtained from the experiment in Tokyo for the method described in Sec. 4.2.3. This section presents results of spot extraction yielded using the proposed method with 64 correct spots in the 2019 version of Walker + cherry blossom viewing in Tokyo. Table 1 presents results found using the method described in Sec. 4.2.3: 1. A spot with many tweets is included in the square is “Contain Gray below"; 2. A spot with many tweets is included in the circle is “Gray within 1 km”; 3. The circle includes locations where the tweets are medium – "White within 1 km".

This result shows few spots for which tweets are concentrated in a square denoted as 1 because the latitude and longitude designated as spots in Walker + are assigned one point. The park has a larger area. Additionally, it is assumed that cherry blossoms in the park are not all at one point. Instead, they are distributed widely throughout the park. Cherry blossoms rarely exist only in a 100 m square area. Therefore, the spot area size can strongly affect results. For this reason, the numbers of spots found within a 1 km circle of 2 or 3 are greater. Some places were not extracted as spots by any determination method, probably because tweets are posted widely from many locations in the park. They might be few compared to those found for the entirety of Tokyo in a relative judgment based on the heat map. Results demonstrate that tweets were few in parks with low visibility. Moreover, extraction using the proposed method was difficult.

4.3.3 Spot Detection Rate in Tokyo

The spot detection rate in Tokyo is presented next. Table 2 presents a comparison between the 64 spots in Sec. 4.3.2 and the extraction results obtained using three determination methods. The spot detection rate is shown as equation (2).

$$\text{Spot Detection Rate} = \frac{\text{Matching}}{\text{Total Spots}} \qquad (2)$$

From these results, it was difficult to ascertain the correct data based on the latitude and longitude of one point, but results suggest that accuracy of 0.8 was obtainable within a 1 km range using the detection method. Setting of the park size range presents issues for future studies, but the capability of extracting major spots using the proposed method was confirmed. This result can engender automation of cherry blossom spot extraction using tweets.

Table 1: Extraction results for Tokyo obtained using the determination method

	Contains Gray below	Gray within 1 km	White within 1 km		Contains Gray below	Gray within 1 km	White within 1 km
Ark hills	0	0	0	Chidorigafuchi park	1	1	1
Asukayama park	1	1	1	Central park	0	0	0
Senzoku pond park	0	0	1	Tokyo midtown	1	1	1
Ikegami honmon temple	1	1	1	Toshimaen sakura festival	0	0	1
Inokashira gift park	1	1	1	Toneri park	0	0	1
Ueno gift park	1	1	1	Toyama park	1	1	1
Ukima park	0	1	1	Sakuragaoka park	0	0	1
Baigan temple	1	1	1	Sayama park	0	0	1
Tamagawadai park	0	1	1	Jindai botanical garden	1	1	1
Lake okutama	0	0	0	Hikarigaoka park	0	1	1
Otonashishinsui park	1	1	1	Yoyogi park	1	1	1
One Green road	0	1	1	Nishiarai park	0	0	1
Kamanofuchi park	0	0	0	Hibiya park	1	1	1
Former iwasaki garden	0	1	1	Hamarikyu gardens	0	1	1
Former shibarikyu garden	1	1	1	Hamura weir	1	1	1
Former furukawa garden	0	1	1	Fujimori park	1	1	1
Kiyosumi garden	0	1	1	Houmyou temple approach	1	1	1
Koishikawa korakuen	0	1	1	Mizumoto park	0	0	0
Koganei park	0	1	1	Myoujinsita park	0	0	0
Showa kinen park	0	0	1	Mukoujima hundred gaeden	0	0	0
Kotta river	0	0	0	Musashin park	0	0	1
Komazawa plympic park	0	1	1	Meijijinguugaien	0	1	1
Sun shine city	1	1	1	Meguro river	1	1	1
Shiotakouchitsutsumi	0	0	0	Roppongi mohri garden	1	1	1
Shiba park	1	1	1	Yaesu sakura street	1	1	1
Shakujii park	0	0	0	Yomiuri land	0	0	0
Shinjuku gyoen	1	1	1	Rikugien	1	1	1
Sumida park	1	1	1	Sotobori park	1	1	1
Sendaiborigawa park	0	0	0	Kinuta park	0	0	1
Zenpukuji river green	0	0	0	Tatsuminomori green park	0	0	1
Takiyama park	0	0	1	Harimazaka	1	1	1
Tama river embankment	0	0	1	Yasukuni shrine	1	1	1

Table 2: Spot detection rate results for Tokyo

	Contains Gray below	Gray within 1km	White within 1km
Matching	27	39	51
Total Spots	64	64	64
Spot Detection Rate	0.42	0.61	0.80

From these results, it was difficult to ascertain the correct data based on the latitude and longitude of one point, but results suggest that accuracy of 0.8 was obtainable within a 1 km range using the detection method. Setting of the park size range presents issues for future studies, but the capability of extracting major spots using the proposed method was confirmed. This result can engender automation of cherry blossom spot extraction using tweets.

4.3.4 Multiple Spots can be Dense

This section presents a description of the heat map results obtained when multiple spots are dense. As an example, Fig. 5 depicts the area around Kita-ku, Tokyo. Spots presented as circles in the figure are cherry blossom spots that were correct answer data. In the figure, A–E respectively stand for Kita-ku Chuo Park, Otonashi Shinsui Park, Asukayama Park, Former Furukawa Garden, and Rikugien Garden.

From Table 1, spots other than Kita-ku Chuo Park in A can be extracted as spots using the proposed method. Regarding A, the tweet exists in the circle, but it was not judged to be a Gray part or a White part because of the small number of tweets compared to those of Tokyo overall.

Additionally, tweets are widely distributed in places other than spots. Of these, tweet origins are most concentrated around stations such as Oji Station, Komagome Station, and Sugamo Station. This concentration is characteristic of areas near stations. Many people can tweet there while waiting for a train or a bus or when loitering around the station when shopping or eating meals. Therefore, when performing an analysis including tweets around the station, the number of excluded spots such as A can be expected to increase.

4.3.5 Time Series Analysis Results

This section presents results of estimation obtained after considering time series for spots that can be detected as cherry blossom spots. As an example, Meguro River results obtained for 2017–2019 are presented in Figs. 6–8. Meguro sightseeing spots can be detected every year, as described in the preceding section. The number of tweets varies every year, but cherry blossom tweets are readily detectable.

Figures 9–11 present results obtained from performing best-time estimation using the proposed method to assess tweets of each year. It tends to blossom at the end of March every year, but each year can produce a different outlook estimate. The accuracy of this best-time estimation result was confirmed using information related to the flowering and full bloom of the Meguro River in 2019, as shown for the Sakura channel of Weathernews as an example [13]. According to the site, the Meguro River in 2019 will bloom on March 21 and will be in full bloom on March 30. The estimation

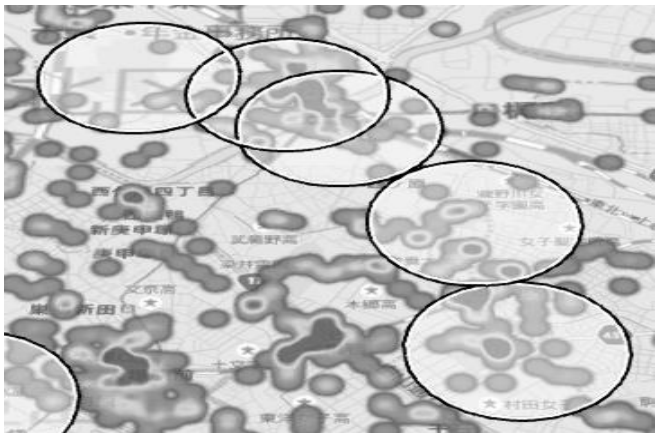

Figure 5: Example of multiple spot judgment results.

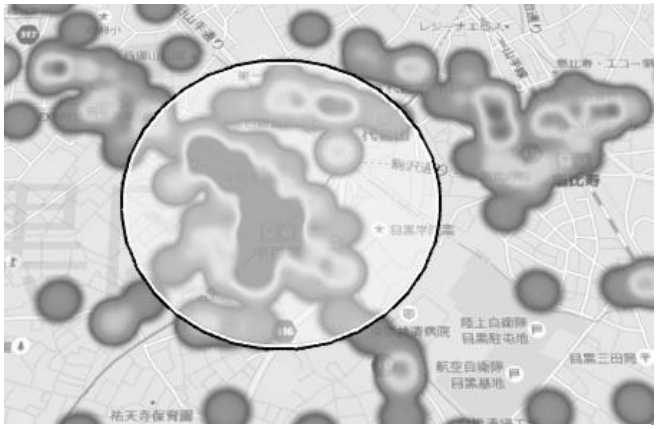

Figure 6: Meguro River judgment results of 2017.

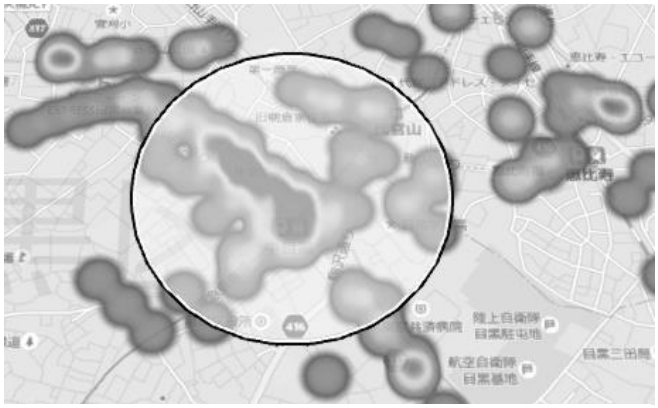

Figure 7: Meguro River judgment results of 2018.

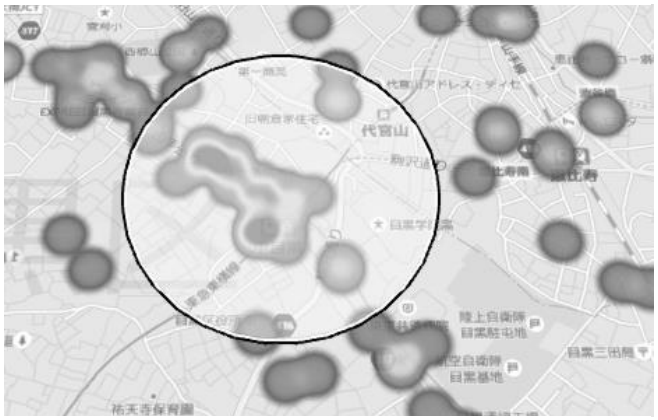

Figure 8: Meguro River judgment results of 2019.

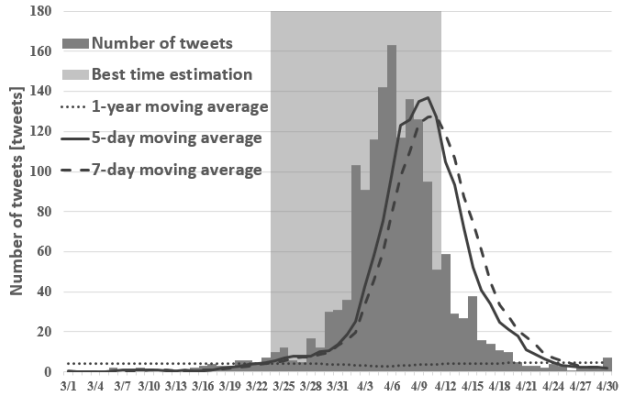

Figure 9: Estimation result of the cherry blossoms in Meguro River in 2017.

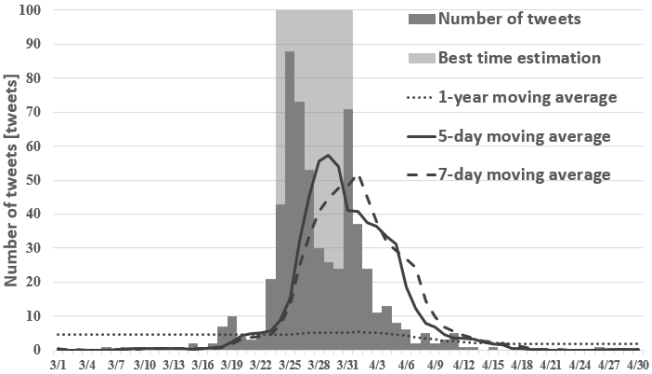

Figure 10: Estimation result of cherry blossoms in Meguro River in 2018.

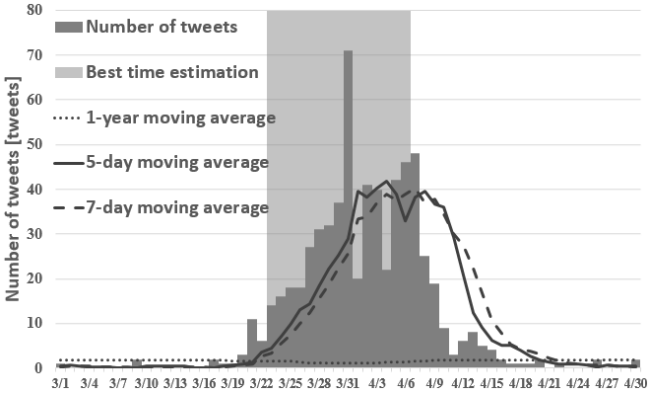

Figure 11: Estimation result of cherry blossoms in Meguro River in 2019.

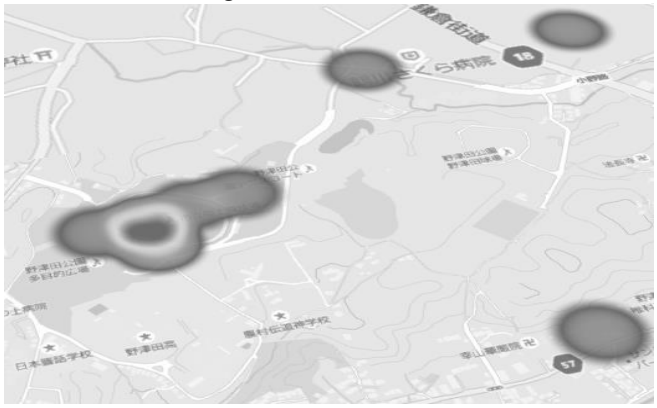
Figure 12: Nozuta Park, newly detected in this experiment

achieved using our proposed method estimates blooming around March 22 – April 1. These results confirmed that the proposed method can estimate flowering to full bloom with sufficient accuracy.

4.3.6 Extracting Spots Other than Correct Answers

Here, places other than those in Table 1, the correct answer spots, are described. After extraction using the heat map, many spots other than the correct spots are extracted in the experiment area of Tokyo.

As described in Sec. 4.3.4, tweets are most concentrated in places where people gather, such as around stations. Next, famous spots are extracted as cherry blossom spots, but not all tourist spots are listed in tourist information magazines and websites. However, extracted locations at which tweets are concentrated might not only be possible for locations that many people know but also locations at which a certain number of tweets are obtainable, such as locally famous locations.

As an example, Fig. 12 presents an example of Nozuta Park. In this figure, no circle signifying a correct answer is displayed, but a visualization result suggesting a correct answer spot was obtained. These results demonstrated that real-time analysis of tweets can automatically extract not only widely known locations but also local cherry blossom viewing spots. The analysis can provide such information. However, when using the proposed method, uniform spot extraction of the experiment range using a heat map might cause errors in local spot extraction because of the influence of tweet amounts around a station. Therefore, improvement of the extraction method is left as a topic for future study.

5 CONCLUSION

As described in this paper, we evaluated a regional analysis method based on location and time-series changes by providing real-time tweets with location information from Twitter. Our existing method provides an estimation of the viewing time for a specific tourist spot. Therefore, the method was constrained to using cases in which travelers have information about the travel area. As described in this paper, as stages 1 and 2, we propose a method for extracting information related to sightseeing spots using geotagged tweets. Subsequently, we demonstrate, experimentally, a method to extract information related to sightseeing spots.

To confirm the usefulness of the proposed method, we conducted experiments demonstrating the automatic extraction of tourist attractions using the proposed method, and experiments for estimating best times using existing methods. As a result, we demonstrated that even a small number of geotagged tweets can be useful to extract spots using location information accumulated over time. Additionally, we showed that the correct spots in Tokyo can be extracted with accuracy of 0.8. Furthermore, we confirmed that the proposed method can extract tourist spots other than the correct answer, such as famous local spots. It was also shown that the existing method is useful for spots extracted using the proposed method to estimate the best time for each sightseeing spot.

These results demonstrate the usefulness of SNS to provide real-time information. Therefore, we were able to demonstrate the possibility of spot extraction using the proposed method, but the scope of its application must be verified further. Moreover, although it is possible to estimate the best time for a spot that can be extracted using the proposed method, we confirmed that the existing methods are not accurate in places where there are tweets of many different types of tweets in areas with high concentrations of people, such as around stations. In future research, it will be necessary to examine a better spot extraction method combining the proposed method described herein with conventional methods and to consider automating tourist spot extraction for real-time analysis.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 18K13254.

REFERENCES

[1] Twitter, It's what's happening, URL < https://Twitter.com/> (2015).
[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, Vol.10, No.3–4, pp.163–173, IARIA (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), pp.425–428, ACM (2014).
[4] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors," pp.851–860, WWW 2010 (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, pp.320–325, ICICCT (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, IEEE (2012).
[7] GeoNLP Project, A place name information processing system which maps sentences automatically, URL < https://geonlp.ex.nii.ac.jp/> (2019).
[8] Twitter Developers, Twitter Developer official site, URL <https://dev.twitter.com/> (2015).
[9] Y. Hashimoto, and M. Oka, "Statistics of Geo-Tagged Tweets in Urban Areas (<Special Issue>Synthesis and Analysis of Massive Data Flow)," Vol.27, No.4, pp.424–431, JSAI (2012) (in Japanese).
[10] National Agriculture and Food Research Organization, Simple reverse geocoding service, URL <http://www.finds.jp/wsdocs/rgeocode/index.html.ja> (2015).
[11] MeCab, Yet Another Part-of-Speech and Morphological Analyzer, URL < http://mecab.google-code.com/svn/trunk/mecab/doc/index.html> (2015).
[12] M. Takahashi, M. Endo, S. Ohno, M. Hirota, and H. Ishikawa, "Automatic detection method of tourist spots using SNS," In Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics (WIMS 2020), pp.91–96, ACM (2020).
[13] WEATHERNEWS Inc., Sakura ch., URL < https://weathernews.jp/s/sakura/spot/4450/> (2019).

(Received January 29, 2019)
(Revised August 21, 2019)

Masaki Endo earned a B.E. degree from Polytechnic University, Tokyo and graduated from the course of Electrical Engineering and Computer Science, Graduate School of Engineering Polytechnic University. He received an M.E. degree from NIAD-UE, Tokyo. He earned a Ph.D. Degree in Engineering from Tokyo Metropolitan University in 2016. He is currently an Associate Professor of Polytechnic University, Tokyo. His research interests include web services and web mining. He is also a member of DBSJ, NPO STI, IPSJ, and IEICE.

Munenori Takahashi is currently studying at Polytechnic university.
His research areas include big data and web mining.

Masaharu Hirota received a Doctor of Informatics degree in 2014 from Shizuoka University. After working for the National Institute of Technology, Oita College, he has been working as an Associate Professor in the Faculty of Informatics, Okayama University of Science since April, 2017. His research interests include photograph, GIS, multimedia, and visualization. He is a member of ACM, DBSJ, and IPSJ.

Makoto Imamura He received a M.E. degree from Kyoto University of Applied Mathematics and Physics in 1986 and a Ph.D. degree from Osaka University of the Information Science and Technology in 2008. During 1986–2016, he worked for Mitsubishi Electric Corp. In April 2016, he moved to the school of Information and Telecommunication Engineering at Tokai University as a Professor. His research interests include machine learning, model-based design and Prognostics and Health Management (PHM).

Hiroshi Ishikawa earned B.S. and Ph.D. degrees in Information Science from The University of Tokyo. After working for Fujitsu Laboratories and becoming a full Professor at Shizuoka University, he became a full Professor at Tokyo Metropolitan University in April, 2013. His research interests include databases, data mining, and social big data. He has published actively in international refereed journals and conferences such as ACM TODS, IEEE TKDE, VLDB, IEEE ICDE, and ACM SIGSPATIAL. He has authored several books: Social Big Data Mining (CRC Press). He is a fellow of IPSJ and IEICE and is a member of ACM and IEEE.