Examination of Best-time Estimation for Each Tourist Spots by Interlinking using Geotagged Tweets

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

Daiju Kato
BI Quality Assurance Division
WingArc1st Inc.
e-mail:kato.d@wingarc.com

Masaharu Hirota
Department of Information Science, Okayama University of Science, Japan
e-mail: hirota@mis.ous.ac.jp

Hiroshi Ishikawa
Graduate School of System Design
Tokyo Metropolitan University, Japan
e-mail: ishikawa-hiroshi@tmu.ac.jp

***Abstract*—Numerous studies have been conducted to analyze social media data in real time and to extract events occurring in the real world. A benefit of analysis using data with position information is that it can accurately extract an event from a target area to be analyzed. However, because data with position information are scarce among all social media data, the amount to analyze is insufficient for almost all areas: we cannot fully extract most events. Therefore, efficient analytical methods must be devised for accurate extraction of events with position information, even in areas with few data. For this study, we estimate the time of biological season observation in areas and sightseeing spots with interlinkage of tweet location information. Herein, we explain the analysis results obtained using information from interpolation and analysis of cherry blossoms in Japan in 2016.**

Keywords–trend estimation; phenological observation; Twitter

I. INTRODUCTION

This paper is an extended version of earlier published work [1]. After improvement of our algorithm, this report describes more accurately estimated best times for viewing organisms at tourist spots using interlinking.

In recent years, because of the wide dissemination and rapid performance improvement of various devices such as smart phones and tablets, diverse and vast data are generated on the web. Particularly, social networking services (SNSs) have become popular because users can post data and various messages easily. Twitter [2], 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.

Sightseeing has come to be regarded as an extremely important growth field in Japan for revival of its powerful economy [3]. Tourism, with its strong economic ripple effects, is expected to produce benefits from regional revitalization and employment opportunities by accommodating world tourism demand, including that from rapidly growing Asia. In addition, people around the world can discover and disseminate the charm of Japan and can promote mutual understanding with other countries.

In addition to the promotion of tourism to Japan, the progress of domestic travel is important. A nation with modern tourism must build a community society that serves regional economies well, attracting tourists widely. Moreover, it is necessary to cultivate tourist areas full of individuality and to promote their charm positively.

According to a survey reported in the Inbound Landing-type Tourism Guide [4] by the Ministry of Economy, Trade and Industry (METI), tourists want real-time information and local unique seasonal information posted on websites. Current websites provide similar information in the form of guidebooks. Nevertheless, the information update frequency of that medium is low. Because each local government, tourism association, and travel company independently provides information about travel destination locales, it is difficult for tourists to collect information for “now” tourist spots. Therefore, providing current, useful, real-world information for travelers by capturing changes of information in accordance with the season and relevant time period of the tourism region is important for the travel industry. As described herein, we define "now" information as information that travelers require for tourism and disaster prevention such as best flower-viewing times, festivals, and local heavy rains. As one might expect, the period estimated for disaster prevention information would be an estimate of the “worst” time instead of the best time.

We consider a method for estimating the best time for tourists to make phenological observations, such as the best time for viewing cherry blossoms and autumn leaves in various regions by particularly addressing phenological observations assumed for “now” in the real world. We define "now" as information for tourism and disaster prevention required by travelers during travel, such as the best flower-viewing times, interesting festivals, and the likelihood of locally heavy rains.

Tourist information for best times requires a peak period, which means that the best time is neither a period after or before falling flowers, but a period to view blooming flowers. Furthermore, the best times differ among regions and locations. Therefore, for each region and location, it is necessary to estimate the best time for phenological observations. Estimating best-time viewing periods requires the collection of large amounts of information having real-time properties. For this study, we use Twitter data obtained from many users throughout Japan. We use Twitter, a typical microblogging service, and also use geotagged tweets that include position information sent in Japan to ascertain the best time (peak period) for biological season observation by region. We proposed a low-cost estimation method [5]. Using this method, prefectures and municipalities showing a certain number of tweets with geotags can be estimated with a relevance rate of about 80% compared to the flowering day / full bloom day of cherry blossoms observed by the Japan Meteorological Agency. The geotagged tweets that are used with this method are useful as social indicators that reflect real-world circumstances. They are a useful resource supporting a real-time regional tourist information system in the tourism field. Therefore, our proposed method might be an effective means of estimating the best time to view events other than biological seasonal observations.

However, to analyze the information of each region from Twitter data, it is necessary to specify the location from tweet information. Because geotagged tweets can identify places, they are effective for analysis. However, because geotagged tweets account for a very small proportion of the total information content of tweets, it is not possible to analyze all regions. For this research, we propose a method of estimating the best time by tourist spot by performing interlinkage using geo-tagged tweets. We conducted experiments to estimate the position around areas not identified by locations based on the amounts of regional information.

The remainder of the paper is organized as follows. Section II presents earlier research related to this topic. Section III describes our proposed method for estimating the best-time of phenological observation through interlinkage using regional amounts. Section IV describes experimentally obtained results obtained using our proposed method, along with a discussion of the results. Section V presents a summary of the contributions and future work.

II. RELATED WORK

The amounts of digital data are expected to increase greatly in the future because of the spread of SNSs. Numerous reports describe studies of the effective use of these large amounts of digital data. Some studies use microblogs to conduct real-world understanding and prediction by analyzing information transmitted from microblogs. Kleinberg [6] describes detection of a "burst" of keywords signaling a rapid increase in time-series data. Sakaki et al. [7] explain their proposal of a method to detect events such as earthquakes and typhoons based on a study estimating real-time events from Twitter. Kaneko et al. [8] propose a method of detecting an event using geotagged non-photo tweets and non-geotagged photo tweets, as well as geotagged photo tweets. Yamagata et al. [9] propose a real-time urban climate monitoring method using geographically tagged tweets, demonstrating the effectiveness of tweets for urban risk management. Consequently, various methods for extracting event and location information have been discussed. Nevertheless, although event detection has been done in earlier studies, no discussion of the event validity period has been reported. As described herein, after proposing a method for extracting such information, we estimate "now" in relation to tourism information, such as the full bloom period of phonological observations. Additionally, we address an important difficulty related to geotagged content analysis: what amount of data is effective for an analyzed area?

Next, along with rising SNS popularity, real-time information has increased. Analysis using real time data has become possible. Many studies have examined efficient methods for analyzing large amounts of digital data. Some studies have been conducted to predict real world phenomena using large amounts of social big data. Phithakkitnukoon et al. [10] analyze details of traveler behavior using data from mobile phone GPS location records such as departure place, destination, and traveling means on a personal level. Mislove et al. [11] develop a system that infers a Twitter user's feelings from the tweet text and visualizes changes of emotion in space–time. After research to detect events such as earthquakes and typhoons, Sakaki et al. [7] propose a method to estimate real-time events from Twitter tweets. Cheng et al. [12] estimate Twitter users' geographical positions at the time of their contributions, without the use of geotags, by devoting attention to the geographical locality of words from text information in Twitter posted articles. Although various studies have analyzed spatiotemporal data, research to estimate the viewing period using interlinkage is a new field.

III. OUR PROPOSED METHOD

This section presents a description of an analytical method for target data collection. It presents best-time estimation to obtain a guide for phenological change from Twitter data in Japan. Our proposal is portrayed in Figure 1.

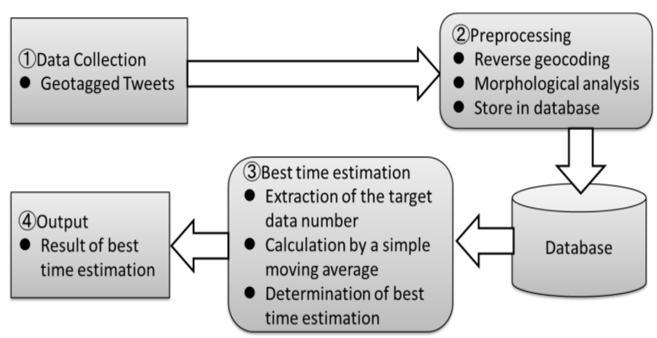

Figure 1. Our proposal.

We describe the best-time estimation method of organisms by analysis using a moving average method for geotagged tweets that include organism names. The best-time estimation in this paper is to estimate the period of time during which the creatures at the tourist spots are useful for sightseeing. Such information can be useful reference information when visiting tourist spots. The information supports estimation of the period during which a tourist can enjoy the four seasons by viewing cherry blossoms and autumn leaves. Hereinafter, Section III.A describes collection of geotagged tweets to be analyzed. Section III.B describes preprocessing for conducting analysis. Section III.C explains the best-time estimation method. In our proposed method up to now, the numbers of geotagged tweets have been small. It is possible to estimate the best time in a prefecture unit or municipality, but finely honed analyses have not been possible. Estimating the best time to visit sightseeing spots with finer granularity is possible using the method with interlinkage proposed in this paper. We propose two improvements of the best estimation method using interlinkage based on the best-time estimation method without interlinkage using the moving average method we have proposed. Section III.C.1 presents interlinkage using Kriging. Section III.C.2 presents interlinking using the average of regional amounts. Section III.D describes the output of the estimation result.

A. Data collection

This section presents a description of the Method of (1) data collection presented in Figure 1. Geotagged tweets sent from Twitter are a collection target. The range of geotagged tweets includes the Japanese archipelago (120.0°E – 154.0°E, and 20.0°N – 47.0°N) as the collection target. The collection of these data was done using a streaming API [13] provided by Twitter Inc.

Next, we describe the number of collected data. According to a report presented by Hashimoto et al. [14], among all tweets originating in Japan, about 0.18% are geotagged tweets: they are rare among all data. However, the geotagged tweets we collected are an average of 500 thousand tweets per day. We use about 250 million geotagged tweets from 2015/2/17 through 2017/5/13. We calculated the best time for flower viewing, as estimated using the processing described in the following sections using these data.

B. Preprocessing

This section presents a description of the method of (2) preprocessing shown in Figure 1. Preprocessing includes reverse geocoding and morphological analysis, as well as database storage for data collected through the processing described in Section III.A.

From latitude and longitude information in the individually collected tweets, reverse geocoding is useful to identify prefectures and municipalities by town name. We use a simple reverse geocoding service [15] available from the National Agriculture and Food Research Organization in this process: e.g., (latitude, longitude) = (35.7384446°N, 139.460910°E) by reverse geocoding becomes (Tokyo, Kodaira City, Ogawanishi-cho 2-chome).

Morphological analysis divides the collected geo-tagged tweet morphemes. We use the “Mecab” morphological analyzer [16]. By way of example, "桜は美しいです” (in English “Cherry blossoms are beautiful.”)" is divided into "(桜 / noun), (は / particle), (美しい / adjective), (です / auxiliary verb), (。 / symbol)".

Preprocessing accomplishes the necessary data storage for best-time viewing, as estimated based on results of the processing of the data collection, reverse geocoding, and morphological analysis. Data used for this study were the tweet ID, tweet post time, tweet text, morphological analysis result, latitude, and longitude.

C. Estimating best-time viewing

This section presents a description of the method of (3) best-time estimation presented in Figure 1. Our method for the best-time viewing processes the target number of extracted data and calculates a simple moving average, yielding an inference of the best time to view the flowers. The method defines a word related to the best-time viewing: the target word. Table I shows that the target word is a word including Chinese characters, hiragana, and katakana, which represents an organism name and a seasonal change.

TABLE I. TARGET WORD EXAMPLES

Items	Target Words	In English
さくら	桜, さくら, サクラ	Cherry blossoms
かえで	楓, かえで, カエデ	Maple
いちょう	銀杏, いちょう, イチョウ	Ginkgo
こうよう	紅葉, 黄葉, こうよう, もみじ, コウヨウ, モミジ	Autumn leaves

A 7-day moving average is based on one week because a tendency exists for tweets to be more numerous on weekends than on weekdays. In addition, the phenological observations which are the current experiment subjects are targeting “events” that happen once a year (e.g., appreciation of cherry blossoms, viewing of autumn leaves, moon viewing). Such events are therefore based on a one-year moving average.

Next, we describe the simple moving average calculation, which uses a moving average of the standard of the best-time viewing judgment. A simple moving average is calculated on a daily basis using aggregate data by the target number of data extraction described above. Figure 2 presents an overview of the simple moving average of the number of days.

We calculate the simple moving average in formula (1) using the number of data going back to the past from the day before the estimated date of the best-time viewing.

$$X(Y) = \frac{P_1 + P_2 + \cdots + P_Y}{Y} \tag{1}$$

$X(Y)$: Y day moving average
P_n : Number of data of n days ago
Y : Calculation target period

The standard lengths of time we used for the simple moving average are 7 days and one year. A 7-day moving average is based on one week as the criterion of the estimated period of full bloom because, as shown in Table II, geotagged tweets of the increases tend to be more numerous on weekends than on weekdays. In addition, the phenological observations which are on the basis of the moving average of best-time the current experiment subjects are targeting "events" that happen once a year (e.g., appreciation of cherry blossoms, viewing estimated in prior years because many such “viewing” events occur every year: cherry blossom viewing, of autumn leaf viewing, and even leaves, moon viewing). Such events are therefore based on a one-year moving average.

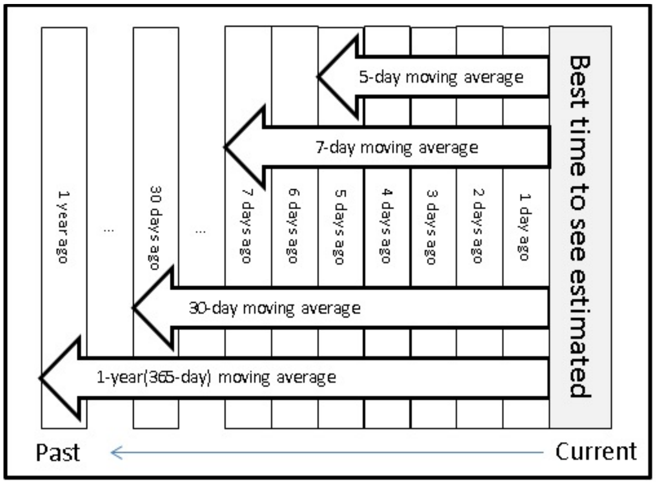

Figure 2. Number of days simple moving average.

TABLE II. TRANSITION OF GEOTAGGED TWEETS (2015/5/9 – 6/3)

Date (Day of the week)	Volume [tweet]	Date (Day of the week)	Volume [tweet]
5/9 (Sat)	117,253	5/22 (Fri)	92,237
5/10 (Sun)	128,654	5/23 (Sat)	55,590
5/11 (Mon)	91,795	5/24 (Sun)	72,243
5/12 (Tue)	87,354	5/25 (Mon)	82,375
5/13 (Wed)	67,016	5/26 (Tue)	83,851
5/14 (Thu)	88,994	5/27 (Wed)	83,825
5/15 (Fri)	89,210	5/28 (Thu)	85,024
5/16 (Sat)	116,600	5/29 (Fri)	121,582
5/17 (Sun)	126,705	5/30 (Sat)	119,387
5/18 (Mon)	89,342	5/31 (Sun)	81,431
5/19 (Tue)	83,695	6/1 (Mon)	76,364
5/20 (Wed)	87,927	6/2 (Tue)	76,699
5/21 (Thu)	86,164	6/3 (Wed)	78,329

Next, we describe a simple moving average of the number of days specified for each organism to compare the 7-day moving average and the one-year moving average. In this study, the best time to view the period varies depending on the specified organism, the individual organism, and the number of days from the biological period.

As an example, we describe cherry blossoms. The Japan Meteorological Agency [17] carries out phenological observations of "Sakura," which yields two output items of the flowering date and the full bloom date observation target. The "Sakura flowering date" [18] is the first day on which blooming of 5–6 or more wheels of flowers occurs on a specimen tree. The "Sakura in full bloom date" is the first day on which about 80% or more of the buds are open in the specimen tree. In addition, "Sakura" is the number of days from general flowering until full bloom: about five days. Therefore, "Sakura" in this study uses a 5-day moving average, which is standard.

Next, we describe an estimated judgment of the best-time for viewing, as calculated using the simple moving average (7-day moving average, one-year moving average, and another biological moving average). It specifies the two conditions as a condition of an estimated decision for the best-time for viewing. Condition 1 is the number of data one day before expression. Formula (2) is a simple moving average, which is greater than that of the estimated best-time to view date. Condition 2 is a case that follows formulas (3) ((A) / (2)) or more. The short number of days by comparison of the 7-day moving average and another biological moving average is A. A long number of days is B.

$$P_1 \geqq X(365) \tag{2}$$
$$X(A) \geqq X(B) \tag{3}$$

Finally, an estimate is produced from conditions 1 and 2. Using the proposed method, a day satisfying both condition 1 and condition 2 is estimated for best-time viewing.

The method presented above is the conventional method we have proposed. However, when interlinkage is not used, it is difficult to estimate the optimum position with fine granularity. Therefore, we propose a method using the following information 1) and 2). Then, interlinkage using 1) or 2) is used to estimate the best viewing time.

1) Interlinkage using Kriging

This section presents the first method of interlinkage, for which we used Kriging [19], an estimation method used for estimating values for points where information was not acquired. We ascertain the distribution of information in the whole space in geostatistics.

Next, the granularity for estimation is shown. For an estimate for Japan as a whole, prefecture units are assumed and acquired by reverse geocoding. However, when conducting more detailed analyses, a difficulty arises: it is impossible to estimate the number of geotagged tweets for each city or town or village or tourist spot. Therefore, based on latitude and longitude information of collected tweets, data are accumulated for each division of land using tertiary mesh data provided by the Land Numerical Information download service of the Ministry of Land, Infrastructure, and Transport [20]. The tertiary mesh is a section of about 1 km square. We attempted estimation through interlinkage using data aggregated for each section of tertiary mesh data. The estimated value of the target data at a certain point S0 is represented in formula (4) as a weighted average of the measured values $Z(S_i)$ (i = 1, 2..., N) at N points S_i existing around point S_0. As described in this paper, we experimentally assigned a +1 weight for 'full bloom' and 'beautiful', and assigned -1 on 'still' or 'falling'. Then we assigned value Z to tweets including the target word and Z. Here, N represents the 30 nearby targeted tweets. λ represents a spherical model that decreases the influence as the distance increases.

$$\hat{z}(S_0) = \sum_{i=1}^{N} \lambda_i Z(S_i) \qquad (4)$$

$Z(S_i)$: Measurement value at i-th position
λ_i : Unknown weighting of measured value at i-th position
S_0 : Predicted position
N : Number of measurements

2) Interlinkage using regional quantity

In this section, we explain the interlinkage method using regional quantities of tweets of city, town, and village units. Conventionally, the best optimum time was estimated using the moving average value without interlinkage using estimation judgment, as described later. As a result, for the analysis of a wide area such as prefecture unit, the R_value can be estimated as about 80%. However, with an estimate of granularity such as by a sightseeing spot, the inability to estimate the viewing period from scarce data is difficult. Therefore, we propose a method of using regional quantities that newly use interlinkage to compensate for the lack of data volume. The proposed method uses the result of reverse geocoding obtained during preprocessing in the previous section. Tweets that were judged as originating from the same municipality by reverse geocoding are summed for each day by city, town, or village. Then, considering the characteristic that the tweets move on a weekly basis, we obtain a 7-day moving average and set the 7-day moving average of the municipalities as the regional quantity of each region. To estimate the best time for viewing, use the value obtained by adding the regional quantity of the municipality where the sightseeing spot is located to the tweet amount of the sightseeing spot to be estimated.

D. Output

This section presents a description of the method of (4) output presented in Figure 1. Output can be visualized using a best-time viewing result, as estimated from processing explained in the previous section. A time-series graph presents the inferred results for best-time viewing. The graph presents the number of data and the date, respectively, on the vertical axis and the horizontal axis. We are striving to develop useful visualization techniques for travelers.

IV. EXPERIMENTS

This section presents a description of an experiment to estimate the best-time viewing for cherry blossoms using the method described in Section III. An experiment is conducted to infer the best time to view flowers for the proposed method described in Section III. Section IV.A describes the dataset used for optimal time reasoning to see flowers in full bloom. As an estimation result by sightseeing spot, Section IV.B presents the estimation result without using interlinkage, and the best estimation result obtained using interlinkage in Section IV.C. Section IV.D presents a comparison of the experiment results from Section IV.B and Section IV.C.

A. Dataset

Datasets used for this experiment were collected using streaming API, as described for data collection in Section III.1. The data, which include about 250 million items, are geotagged tweets from Japan during 2015/2/17 – 2017/5/13.

The estimation experiment conducted to ascertain the best-time viewing of cherry blossoms uses the target word "cherry blossom," which is "桜" and "さくら" and "サクラ" in Japanese. We analyzed tweet texts that included the target word. About 100,000 tweets during in the experiment period included the subject word.

The subject of the experiment was set as tourist spots in Tokyo. In this report, we describe "Takao Mountain," "Showa Memorial Park," "Shinjuku gyoen," and "Rikugien." Figure 3 presents the target area locations. A, B, C, and D in the figure respectively denote "Takao Mountain," "Showa Memorial Park," "Rikugien," and "Shinjuku Gyoen." A and B are separated by about 16 km straight-line distance. B and C are about 32 km apart. C and D are about 6 km apart. In this experiment, about 30,000 tweets including the target word in Tokyo were found. In this experiment, all tweets made by the same user are also used as analysis targets if they are tweets including the target word.

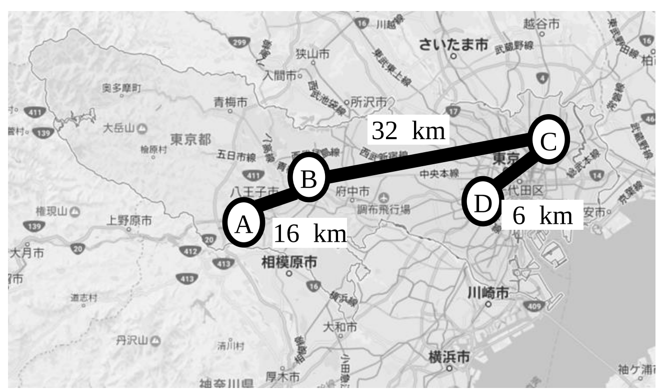

Figure 3. Position of the target area.

We use these datasets to estimate the optimum time for the sightseeing spots in Tokyo using experiments without interlinkage. We conducted experiments of the following two kinds. The first is an experiment using the number of tweets including the target word and the sightseeing spot name without interlinkage. This experiment uses the conventional method we proposed. This experiment was compared as Baseline to confirm the usefulness of interlinkage proposed in this paper. The second is an experiment using interlinkage. In this experiment, we used two methods: interlinkage using Kriging of Section III.C.1) and interlinkage using regional quantity described in Section III.C.2).

B. Estimation experiment for best-time viewing without interlinkage

This section presents experimentally obtained results for estimating the best time without using interlinkage from tweets containing a target word and sightseeing spot name. Figure 4 presents results for the estimated best-time viewing in 2016 using the target word ‘cherry blossoms’ in the target tourist spots. The dark gray bar in the figure represents the number of tweets. The light gray part represents best-time viewing as determined using the proposed method. Additionally, the solid line shows the 5-day moving average. The dashed line shows the 7-day moving average. The dotted line shows the one-year moving average.

At tourist spots targeted for the experiment in 2016, as portrayed in Figure 4, many data were obtained for C and D. The maximum number of tweets per day was about 30. These results confirmed that some estimation can be accomplished using near-site estimation method without interpolation. However, best-time viewing cannot be done in A and B because of the very small number of tweets.

This difficulty applies to many sightseeing spots in Japan. In fact, a difficulty exists by which it is impossible to estimate the best-time to see the sightseeing spots and other fine-grain sight by simply using the number of geotagged tweets. This experimentally obtained result clarified that the method we proposed previously cannot be predictive for detailed areas such as sightseeing spots. This result is attributable to the lack of information volume.

C. Estimation experiment for best-time viewing with interlinkage

In this section, the results of interlinkage using Kriging of Section III.C.1, which is the method proposed in this paper, and interlinkage using the region quantities of Section III.C.2 are shown in Section IV.C.1 and Section IV.C.2.

1) Estimation experiment for best-time viewing by interlinkage using Kriging

Figure 5 portrays an experimentally obtained result from interlinking results for a tertiary mesh including the tourist spots we examined. The notation is the same as that presented in Figure 4.

Apparently, A and B were able to produce an estimate using the proposed method by increasing the number of tweets using interlinkage with surrounding tweets. For C and D, there are days when it can be determined more accurately by interpolating the number of tweets. However, because tweets of negative judgments exist such as "still" or "scattered" among surrounding tweets, in some cases, interpolation excluded the day determined as the best time in Experiment 1. Therefore, the judgment condition of the tweet is subject to further study. These results confirmed the possibility of estimating the peak period, even for an area without tweets, using data interpolation and overall tweet number interpolation.

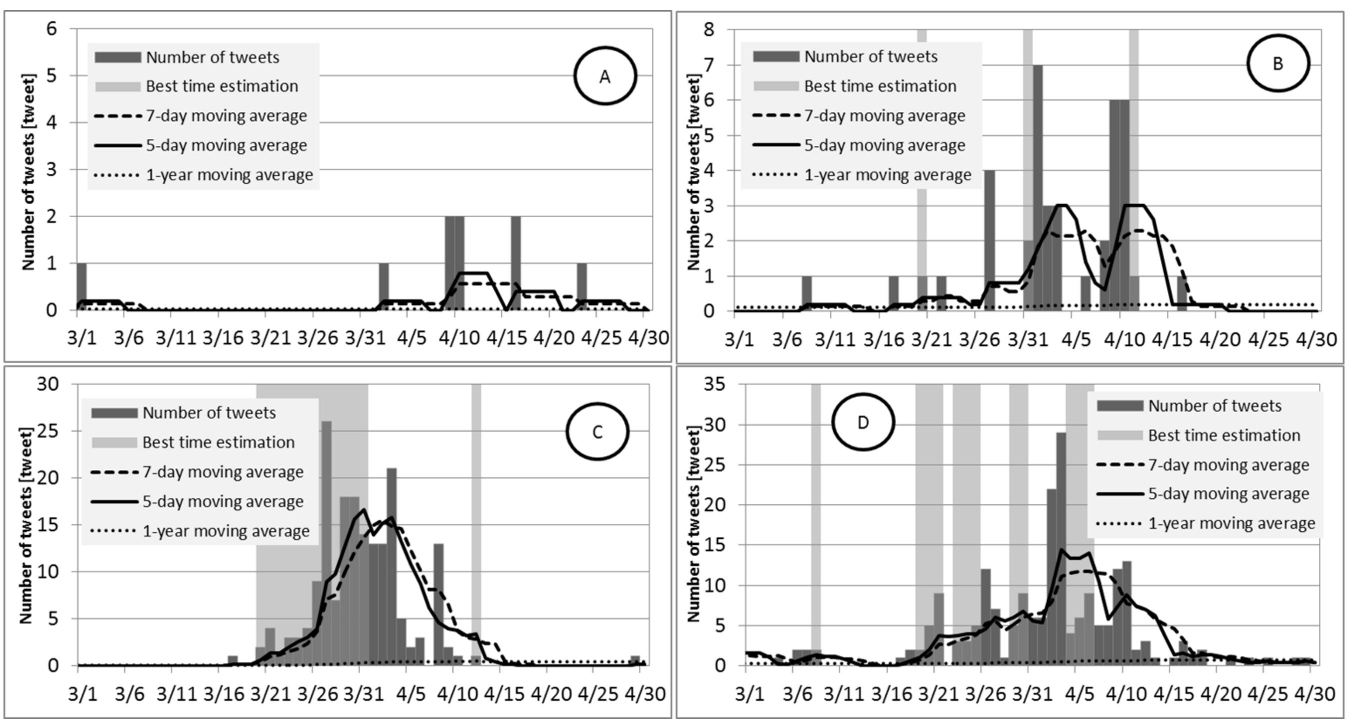

Figure 4. Experimental results obtained using tweets including the target word and the tourist spot name.

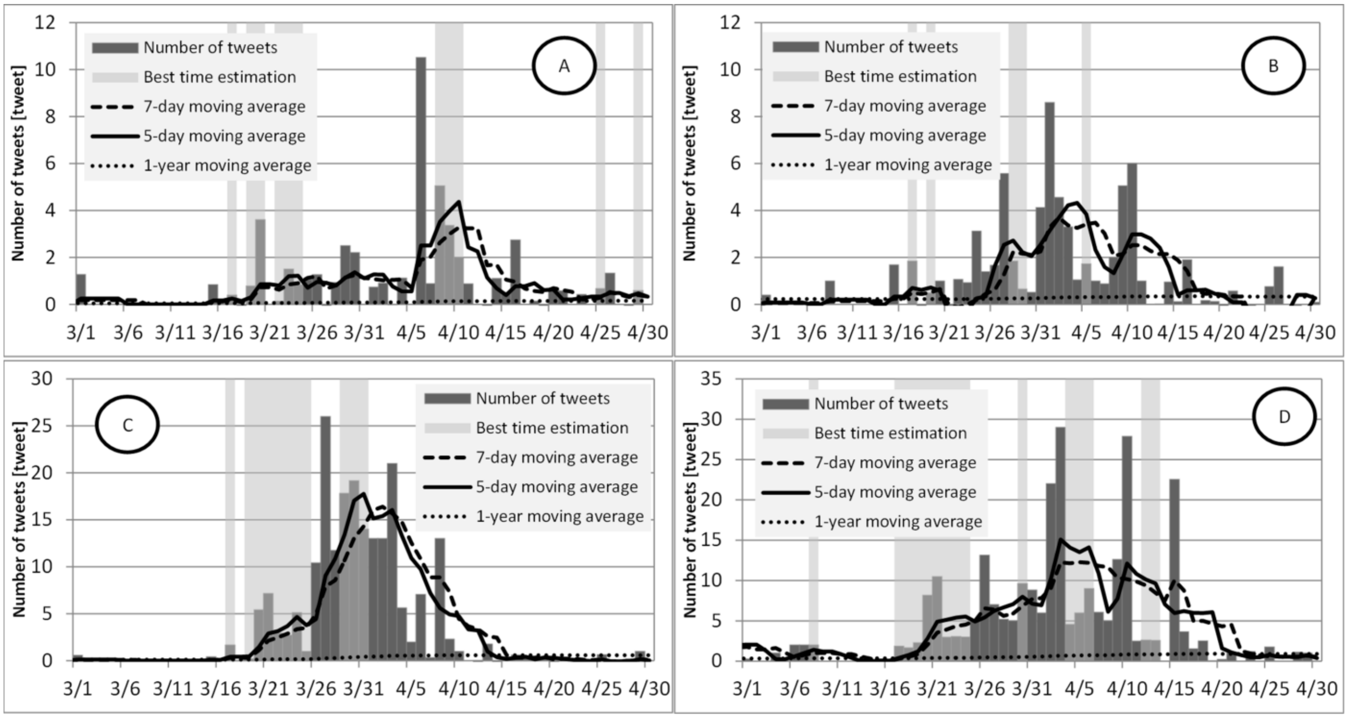

Figure 5. Experimental results obtained using interlinking.

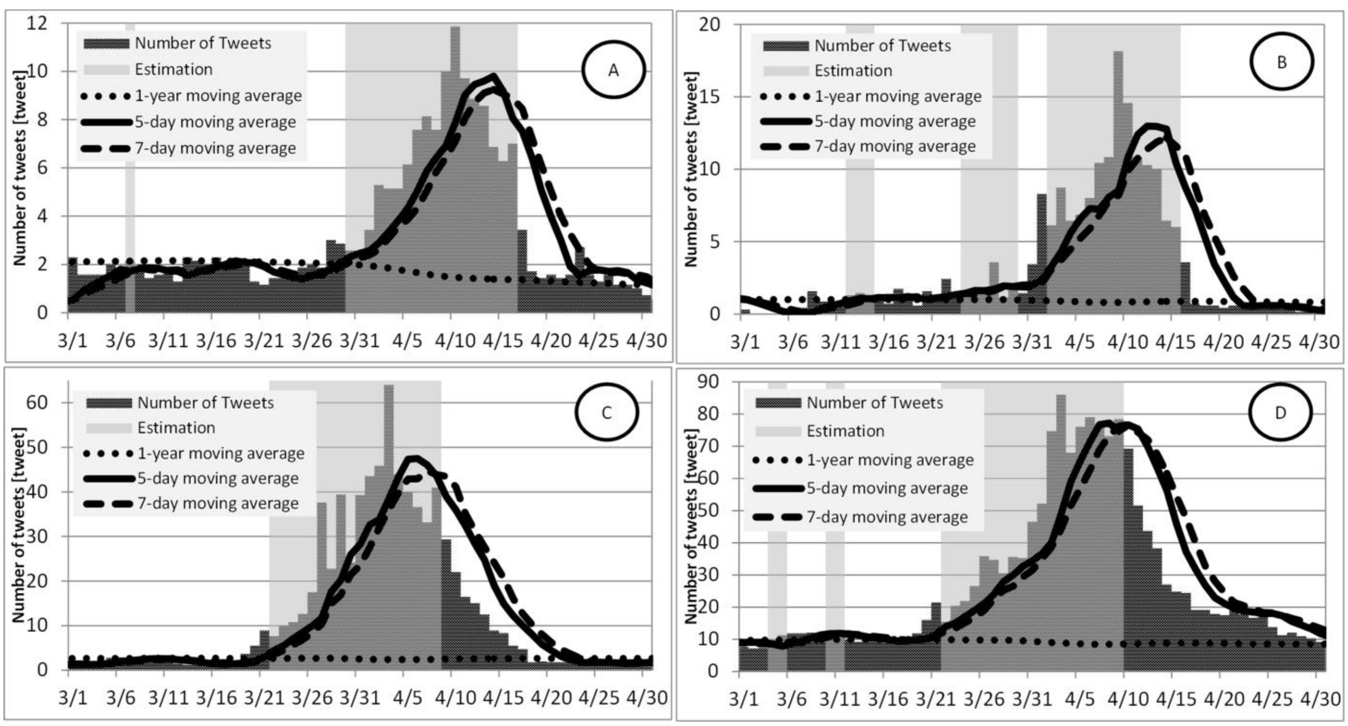

Figure 6. Experimental results obtained using interlinking.

2) Estimation experiment for best-time viewing by interlinkage using regional quantities

In this section, we explain experimentally obtained estimation results obtained using interlinkage of regional quantities. Figure 6 presents results obtained using interlinking. The notation is the same as the notation used for Figure 3 in the previous section.

Apparently, A and B can produce an estimate using the proposed method by increasing the number of tweets using interlinkage with surrounding tweets. For C and D, there are days that can be determined more accurately by interpolating the number of tweets. These results demonstrate the possibility of first resolving the difficulty of insufficient information when using sightseeing spot tweet data of the tourist spot area along with interpolation, and then estimating the peak period for each tourist spot.

Furthermore, even given the same area of Tokyo, the times estimated for A and B are later than those of C and D. These later results is attributable to the fact that A and B are at higher altitudes than either C or D. Flowers can be expected to bloom later in the year there.

Therefore, using interlinkage, one can confirm differences in surroundings even within one prefecture. More detailed analysis becomes possible with interlinkage using the proposed method.

D. *Comparing estimations and observed data for best times for viewing*

This section presents a comparison between experimentally obtained results and observation data. An estimation result obtained without using interlinkage was taken as the Baseline. Then, the experiment using interlinkage with Kriging of Section IV.C.1 was taken as experiment 1. Experiment 2 was based on interlinkage using the regional quantity of Section IV.C.2.

1) Comparison of observation data with best time of viewing estimation using interlinkage with Kriging

Table III presents results of the optimal time for viewing in 2016, as estimated using Kriging. The baseline examination used co-occurring words in tweets including the sightseeing spot name coexisting with the target word "Sakura." Experiment 1 used interlinkage on a tertiary mesh including sightseeing spots. The numerical values in the table are the numbers of tweets including the target word and co-occurrence word in the Baseline. Experiment 1 uses the sum of the number of tweets in the Baseline plus the numerical values by obtained using interpolation. The light gray area shows the date when the fullness prediction was made using the proposed method.

Confirming the flowering day and full bloom period of each sightseeing spot using JMA data is difficult, but this experiment evaluating SNS data for flowering is valid also for use by weather forecasting companies [21] and public service organizations [22] to ascertain the optimum times for viewing based on services and blogs that are used. Arrows indicating the flowering time can be checked manually at tourist sites. Recall and precision using the observed data and the best time to view estimated results are calculated respectively for target areas for 2016 for 3/1 – 4/30 using formula (5) and formula (6). We used the R_value as the recall rate using the proposed method, and the P_value and R_value, respectively, as the relevance rate and recall rate when using the proposed method.

$$P\_value = \frac{Number\ of\ days\ to\ match\ the\ observed\ data}{Number\ of\ days\ in\ best\ time\ to\ see\ estimated} \quad (5)$$

$$R\_value = \frac{Number\ of\ days\ to\ match\ the\ observed\ data}{Number\ of\ days\ of\ observation\ data} \quad (6)$$

TABLE III. COMPARISON RESULTS FOR BASELINE AND EXPERIMENT 1

	Takao mountain		Showa Memorial park		Rikugien		Shinjuku gyoen	
	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2
3/1	1	1.27	0	0.40	0	0.58	0	0.28
3/2	0	0.00	0	0.00	0	0.00	0	0.00
3/3	0	0.00	0	0.00	0	0.00	0	0.00
3/4	0	0.00	0	0.00	0	0.00	1	1.00
3/5	0	0.00	0	0.00	0	0.00	0	0.00
3/6	0	0.00	0	0.00	0	0.00	2	2.00
3/7	0	0.00	0	0.00	0	0.00	2	2.00
3/8	0	0.00	1	1.00	0	0.00	2	2
3/9	0	0.00	0	0.00	0	0.00	0	0.00
3/10	0	0.00	0	0.00	0	0.00	0	0.00
3/11	0	0.00	0	0.00	0	0.00	0	0.00
3/12	0	0.00	0	0.00	0	0.00	0	0.00
3/13	0	0.00	0	0.00	0	0.00	0	0.00
3/14	0	0.00	0	0.00	0	0.00	0	0.00
3/15	0	0.84	0	1.68	0	0.41	0	0.24
3/16	0	0.00	0	0.00	0	0.00	0	0.00
3/17	0	0.40	1	1.85	1	1.67	1	1.85
3/18	0	−0.51	0	−0.43	0	0.02	2	1.64
3/19	0	0.78	0	0.49	0	0.22	2	2.30
3/20	0	3.61	1	1.00	2	5.42	5	8.18
3/21	0	0.00	0	−4.81	4	7.16	9	10.49
3/22	0	0.14	1	1.06	1	2.93	0	2.97
3/23	0	1.51	0	0.93	3	2.70	3	3.06
3/24	0	0.84	0	3.13	3	5.11	3	3.00
3/25	0	1.06	0	1.40	4	1.00	5	5.00
3/26	0	1.27	0	1.67	9	10.41	12	13.14
3/27	0	−0.08	4	5.57	26	26.00	7	7.00
3/28	0	0.94	0	1.85	7	11.74	1	5.19
3/29	0	2.50	0	0.66	18	17.83	5	5.00
3/30	0	2.21	0	0.52	18	19.18	9	9.62
3/31	0	0.00	2	4.13	14	14.00	6	8.82
4/1	0	0.74	7	8.60	13	13.00	6	6.00
4/2	1	0.91	3	4.56	13	13.00	22	22.00
4/3	0	0.00	3	3.30	21	21.00	29	29.00
4/4	0	1.12	0	1.05	5	5.62	4	4.55
4/5	0	0.00	0	1.73	2	2.00	6	6.00
4/6	0	10.52	1	1.00	3	7.05	9	9.00
4/7	0	0.89	0	0.88	0	0.33	5	6.06
4/8	0	5.05	2	2.00	13	13.00	5	5.00
4/9	2	3.37	6	5.05	2	2.29	12	12.62
4/10	2	2.00	6	6.00	1	1.00	13	27.88
4/11	0	0.88	1	1.00	0	0.00	2	2.47
4/12	0	0.00	0	0.00	1	−0.24	3	2.61
4/13	0	−0.50	0	0.02	0	1.79	1	2.55
4/14	0	1.11	0	0.95	0	−0.67	0	−0.37
4/15	0	0.51	0	0.12	0	0.00	1	22.54
4/16	2	2.75	1	1.91	0	0.60	3	3.63
4/17	0	0.00	0	0.00	0	0.42	1	1.54
4/18	0	0.14	0	0.17	0	0.36	2	2.51
4/19	0	−0.34	0	0.13	0	0.30	0	0.07
4/20	0	0.66	0	−0.43	0	−0.20	0	0.21
4/21	0	0.58	0	0.58	0	0.00	1	1.00
4/22	0	0.00	0	0.00	0	0.00	0	0.00
4/23	1	0.43	0	−4.08	0	0.00	1	1.19
4/24	0	0.00	0	0.00	0	0.00	0	0.00
4/25	0	0.68	0	0.76	0	0.64	1	1.75
4/26	0	1.33	0	1.61	0	−1.03	0	−0.26
4/27	0	0.00	0	0.00	0	0.00	0	0.00
4/28	0	−0.25	0	−0.34	0	0.19	1	1.11
4/29	0	0.60	0	0.00	1	1.00	1	1.00
4/30	0	0.00	0	0.09	0	0.00	0	0.00
P_value	0.26	0.38	0.72	0.75	0.89	0.89	0.84	0.75
R_value	0.00	0.20	0.11	0.22	0.67	0.61	0.56	0.50

Experimental results confirmed the tendency by which the relevance ratio and the R_value become higher for tourist spots with fewer tweets. Therefore, we can present the possibility of estimating sightseeing sites with few tweets using interlinkage. However, because the interpolation information amount is insufficient in the current method, it is necessary to improve the interlinking method further. Furthermore, because sightseeing spots with many tweets are affected by tweets of negative judgments in the surroundings, the accuracy is lower than in the Baseline. However, one might be able to estimate more details, such as the start time, using interpolation.

2) Comparison of observation data with best time of viewing estimation using interlinking with regional quantities

Table IV presents results of the optimal time for viewing in 2016, as estimated using regional quantities. Experimentally obtained results confirmed the tendency by which the relevance ratio and the R_value became higher in Experiment 2 than in the Baseline. In addition, A and B, which are at higher altitudes than either C or D, demonstrated regional features: the best viewing time occurs later. These results confirmed the usefulness of the proposed method for best-time estimation for sightseeing spots using interlinkage along with regional data. Using this proposed method, information can be interpolated according to the sightseeing spot. One can obtain better estimation with finer granularity than that available with the conventional method.

Additionally, comparison with Experiment 1 of Section IV.D.1 is good because, in Experiment 1, the tweet contents are analyzed and interpolation is performed using tweets including specific information. Therefore, the tweet number used for interlinkage is less than that in Experiment 2. By contrast, in Experiment 2, the moving average value using all tweets included in the same area judged as the same municipality by reverse geocoding is used without analyzing the tweet contents. Experiment 2 is useful at the present stage, but it seems that the possibility exists that improving Experiment 1 to analyze tweet contents might improve the estimation accuracy.

V. CONCLUSION

As described herein, to improve best-time estimation accuracy and thereby enhance tourist information related to phenological observation, we proposed an interlinking method.

For the first proposed method, information was interpolated using neighbor-weighted tweets on a tertiary mesh including sightseeing spots, thereby indicating the optimum time to view flowers at sightseeing spots. The results of cherry blossom experiments conducted at sightseeing spots in Tokyo in 2016 confirm the tendency for improvement of the estimation accuracy using interlinking. The proposed method using interlinkage for tweets related to organism names might improve the accuracy of estimating the best time in the real world. We confirmed the

TABLE IV. COMPARISON RESULTS FOR BASELINE AND EXPERIMENT 2

	Takao Mountain		Showa memorial park		Rikugien		Shinjuku gyoen	
	Baseline	Exp.2	Baseline	Exp.2	Baseline	Exp.2	Baseline	Exp.2
3/1	1	2.29	0	0.40	0	0.58	0	7.57
3/2	0	1.57	0	0.00	0	0.00	0	7.14
3/3	0	1.57	0	0.00	0	0.00	0	7.29
3/4	0	1.57	0	0.00	0	0.00	1	9.86
3/5	0	2.00	0	0.00	0	0.00	0	11.29
3/6	0	1.86	0	0.00	0	0.00	2	11.86
3/7	0	2.14	0	0.00	0	0.00	2	11.86
3/8	0	1.71	1	1.00	0	0.00	2	12
3/9	0	1.43	0	0.00	0	0.00	0	11.86
3/10	0	1.57	0	0.00	0	0.00	0	11.71
3/11	0	1.71	0	0.00	0	0.00	0	10.71
3/12	0	1.29	0	0.00	0	0.00	0	9.86
3/13	0	2.29	0	0.00	0	0.00	0	9.14
3/14	0	2.14	0	0.00	0	0.00	0	10.14
3/15	0	2.00	0	1.68	0	0.41	0	9.43
3/16	0	2.29	0	0.00	0	0.00	0	8.86
3/17	0	2.00	1	1.85	1	1.67	1	10.14
3/18	0	2.00	0	1.00	0	1.86	2	10.57
3/19	0	2.00	0	0.57	0	3.57	2	11.86
3/20	0	1.29	1	1.57	2	5.43	5	16.00
3/21	0	1.14	0	1.14	4	8.86	9	21.43
3/22	0	1.43	1	2.43	1	7.57	0	15.43
3/23	0	1.43	0	1.43	3	10.00	3	20.43
3/24	0	1.71	0	1.57	3	10.71	3	21.86
3/25	0	1.86	0	1.43	4	12.57	5	26.57
3/26	0	2.00	0	1.71	9	17.43	12	35.86
3/27	0	1.86	4	3.57	26	37.57	7	34.71
3/28	0	3.00	0	1.14	7	22.71	1	30.57
3/29	0	2.86	0	2.14	18	39.43	5	35.57
3/30	0	2.57	0	1.43	18	24.14	9	35.29
3/31	0	2.57	2	3.43	14	39.29	6	46.57
4/1	0	3.43	7	8.29	13	43.57	6	52.14
4/2	1	5.29	3	6.14	13	45.86	22	74.71
4/3	0	5.14	3	8.71	21	64.00	29	86.00
4/4	0	5.14	0	6.43	5	44.00	4	68.00
4/5	0	6.14	0	6.86	2	40.00	6	76.00
4/6	0	7.57	1	8.00	3	36.57	9	79.00
4/7	0	8.14	0	10.43	0	33.14	5	76.86
4/8	0	7.57	2	10.86	13	41.00	5	73.14
4/9	2	10.00	6	18.14	2	29.29	12	78.43
4/10	2	11.86	6	14.57	1	22.00	13	69.29
4/11	0	9.71	1	10.86	0	16.43	2	51.57
4/12	0	8.86	0	10.29	1	15.00	3	43.71
4/13	0	8.57	0	10.00	0	12.43	1	38.29
4/14	0	6.86	0	6.43	0	8.86	0	27.00
4/15	0	6.29	0	6.00	0	8.29	1	24.86
4/16	2	7.00	1	3.57	0	5.43	3	24.43
4/17	0	3.43	0	0.71	0	4.71	1	19.14
4/18	0	1.71	0	0.57	0	1.71	2	19.14
4/19	0	1.43	0	0.57	0	1.86	0	17.86
4/20	0	1.57	0	0.43	0	1.71	0	17.57
4/21	0	1.43	0	0.57	0	1.71	1	20.00
4/22	0	1.57	0	0.57	0	1.71	0	17.71
4/23	1	2.71	0	-4.08	0	0.00	1	16.71
4/24	0	1.57	0	0.00	0	0.00	0	18.14
4/25	0	1.29	0	0.76	0	0.64	1	13.86
4/26	0	1.57	0	1.61	0	-1.03	0	11.29
4/27	0	1.14	0	0.00	0	0.00	0	12.00
4/28	0	1.14	0	-0.34	0	0.19	1	11.00
4/29	0	1.00	0	0.00	1	1.00	1	10.29
4/30	0	0.71	0	0.09	0	0.00	0	9.14
P_value	0.26	0.77	0.72	0.74	0.89	0.84	0.84	0.82
R_value	0.00	0.58	0.11	0.44	0.67	0.58	0.56	0.83

possibility of applying this proposed method to estimation of the viewpoint and line of sight in areas and sightseeing spots with few tweets and little location information. However, analysis of the contents of tweets is an experimentally obtained result obtained using few words. Effective interlinking cannot be established. Therefore, future research with manual experimental weighting and geotagged tweets is expected to facilitate further improvements to overcome insufficiencies in the measured values used for interpolation. Additionally, we expect to reconsider the viewing angle estimation conditions.

The second proposed method showed optimal times to see flowers at sightseeing spots by interpolating information using the 7-day moving average of the number of tweets of municipalities, including those of sightseeing spots. This method can estimate the best time for sightseeing spots with fine granularity, yielding predictions in units required for sightseeing.

The results of cherry blossom experiments conducted for tourist spots in Tokyo in 2016 using the proposed method confirmed improvement of the estimation accuracy when using interlinking. The proposed method using interlinkage for tweets related to named organisms (sakura trees) might improve the real-world accuracy of estimating the best times. We confirmed the possibility of applying this proposed method to estimation of viewpoints and lines of sight in areas and sightseeing spots with few tweets and little location information. The proposed method interpolated information and yielded highly accurate estimation, perhaps because of the fact that cherry blossoms bloom with short-term changes and because public interest is high. Therefore, future studies can be conducted to verify whether similar results are obtainable using other biological season observations.

Future studies must also assess the automatic extraction of target words and methods to make future predictions in real time. Eventually, this system might be extended to a system for travelers to obtain travel-destination-related event information and disaster information in real time.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Nos. 16K00157 and 16K16158, and by a Tokyo Metropolitan University Grant-in-Aid for Research on Priority Areas "Research on Social Big Data."

REFERENCES

[1] M. Endo, S. Ohno, M. Hirota, Y. Shoji, and H. Ishikawa, "Examination of Best-time Estimation using Interpolation for Geotagged Tweets," MMEDIA 2017 Proceeding of The Ninth International Conference on Advances in Multimedia, pp. 38-43, 2017.

[2] Twitter. *It's what's happening.* [Online]. Available from: https://Twitter.com/ [retrieved: September 2, 2017]

[3] Japan Tourism Agency. *Tourism Nation Promotion Basic Law*. [Online]. Available from: http://www.mlit.go.jp/kankocho/en/kankorikkoku/index.html [retrieved: April 14, 2017]

[4] Ministry of Economy, Trade and Industry. *Inbound Landing-Type Tourism Guide*. [Online]. Available from: http://www.mlit.go.jp/common/001091713.pdf [retrieved: November 10, 2017] (in Japanese).

[5] M. Endo, Y. Shoji, M. Hirota, S. Ohno, and H. Ishikawa, "On best time estimation method for phenological observations using geotagged tweets," International Workshop on Informatics 2016 (IWIN2016), 2016.

[6] J. Kleinberg, "Bursty and hierarchical structure in stream," In Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1–25, 2002.

[7] 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.

[8] T. Kaneko and K. Yanai, "Visual Event Mining from the Twitter Stream," WWW '16 Companion Proceedings of the 25th International Conference Companion on World Wide Web, pp. 51–52, 2016.

[9] Y. Yamagata, D. Murakami, G. W. Peters, and T. Matsui, "A spatiotemporal analysis of participatory sensing data "tweets" and extreme climate events toward real-time urban risk management," arXiv preprint arXiv:1505.06188, pp. 1–34, 2015.

[10] S. Phithakkitnukoon, T. Teerayut Horanont, A. Witayangkurn, R. Siri, Y. Sekimoto, and R. Shibasaki, "Understanding tourist behavior using large-scale mobile sensing approach: A case study of mobile phone users in Japan," Pervasive and Mobile Computing, 2014.

[11] A. Mislove, S. Lehmann, Y.Y. Shn, and J. Kleinberg, "Bursty and hierarchical structure in stream," In P. Onnela, and Rosenquist, Understanding the Eighth ACM SIGKDD Demographics of Twitter Users, Proceedings. Fifth International AAAI Conference on Knowledge Discovery, Weblogs and Data Mining, Social Media (ICWSM'11), pp. 1–25, 2002. 133-140, 2011.

[12] Z. Cheng, J. Caverlee, and K. Yanai, "Visual Event Mining from the Twitter Stream," WWW '16 CompanionLee, You are where you tweet: a content-based approach to geo-locating twitter users, Proceedings of the 25th 19th ACM International Conference Companion on World Wide Web, pp. 51–52, 2016.on Information and Knowledge Management, 2010.

[13] Twitter Developers. *Twitter Developer official site*. [Online] Available from: https://dev.twitter.com/ [retrieved: April 2, 2017]

[14] 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).

[15] National Agriculture and Food Research Organization. *Simple reverse geocoding service*. [Online]. Available from: https://www.finds.jp/rgeocode/index.html.ja [retrieved: November 18, 2017]

[16] MeCab. *Yet Another Part-of-Speech and Morphological Analyzer*. [Online]. Available from: http://taku910.github.io/mecab/ [retrieved: November 10, 2017]

[17] Japan Meteorological Agency. *Disaster prevention information XML format providing information page*. [Online]. Available from: http://xml.kishou.go.jp/ [retrieved: April 5, 2015]

[18] Japan Meteorological Agency. *Observation of Sakura*. [Online]. Available from: http://www.data.jma.go.jp/sakura/data/sakura2012.pdf [retrieved: April 5, 2015]

[19] M. A. Oliver, "Kriging: A Method of Interpolation for Geographical Information Systems," International Journal of Geographic Information Systems vol. 4, pp. 313–332, 1990.

[20] Ministry of Land, Infrastructure and Transport. *Land Numerical Information download service*. [Online]. Available from: http://nlftp.mlit.go.jp/ksj-e/index.html [retrieved: April 5, 2015]

[21] Weathernews Inc. *Sakura information*. [Online]. Available from: http://weathernews.jp/sakura [retrieved: May 10, 2017]

[22] Japan Travel and Tourism Association. *Whole country cherry trees*. [Online]. Available from: http://www.nihon-kankou.or.jp/sakura/ [retrieved: November 14, 2017]