Regular Paper

Best-Time Estimation Method using Information Interpolation for Sightseeing Spots

Masaki Endo*, Masaharu Hirota**, Shigeyoshi Ohno*, and Hiroshi Ishikawa***

*Division of Core Manufacturing, Polytechnic University, Japan
{endou, ohno}@uitec.ac.jp
**Department of Information Science, Okayama University of Science, Japan
hirota@mis.ous.ac.jp
***Graduate School of System Design, Tokyo Metropolitan University, Japan
ishikawa-hiroshi@tmu.ac.jp

Abstract – With the spread of SNS, many data are transmitted in real time. Some data with position information are included in these data. A benefit of analysis using data with position information is that they can extract an event accurately 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 in almost all areas. In other words, most events cannot be fully extracted. 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 particular areas and sightseeing spots by information interpolation using tweet location information. Herein, we explain the analysis results obtained using interpolation of information related to cherry blossoms and autumn leaves as an example.

Keywords: information interpolation, phenological observation, trend estimation, Twitter

1 INTRODUCTION

In recent years, sightseeing has come to be regarded as an extremely important growth field to revive Japan's powerful economy [1]. Tourism, with its strong economic ripple effect, is expected to benefit regional revitalization and employment opportunities through accommodation of 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 among countries.

In addition to the promotion of tourism to Japan, the progress of domestic travel is important. It is necessary for a nation with modern tourism to build a community society by which regional economies are well-served, attracting tourists widely. Moreover, it is necessary to cultivate tourist areas full of individuality and to promote their charm positively.

According to a survey study of information technology (IT) tourism and services to attract customers [2] 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, information of that medium is not frequently updated. Because each local government, tourism association, and travel company independently provides information about local travel destinations, it is difficult for tourists to collect information for "now" tourist spots. Therefore, the travel industry demands that current, useful, real-world information be provided for travelers by capturing the change of information in accordance with the season and time zone of the tourism region.

We consider a method to estimate the best time for phenological observations for tourism such as the best time for viewing cherry blossoms and autumn leaves in each region by particularly addressing phenology observations assumed for "now" in the real world. We define "now" information as that intended for tourism and disaster prevention required by travelers during travel, such as best flower-viewing times, festivals, and locally heavy rains.

Tourist information for best times requires a peak period: the best time is not a period after or before falling flowers, but a period that is best to view blooming flowers. Furthermore, the best times differ among regions and locations. Therefore, it is necessary to estimate the best time of phenological observation for particular regions and locations. Estimating best-time viewing requires collection of large amounts of information with real-time properties.

For this study, we use Twitter data obtained from many users throughout Japan. Twitter [3], a typical microblogging service, has some geotagged tweets that include position information sent in Japan. We use the data to ascertain the best time (peak period) in biological season observation by region. We proposed a low-cost estimation method [4] by which 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 the real world situation. 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.

Nevertheless, geotagged tweets are extremely few among all tweets. Therefore, a difficulty exists that the data are insufficient for analysis with finer granularity. For this reason, it is necessary to improve the method of interpolating the information of geotagged tweets to conduct further detailed

analyses in areas such as sightseeing spots. For this research, we propose a method of estimating the best time for a particular tourist spot by performing information interpolation based on amounts of regional information. This paper presents results of verification by experimentation using cherry blossoms and autumn leaves.

The remainder of the paper is organized as follows. Chapter 2 presents earlier research related to this topic. Chapter 3 describes our proposed method for estimating the best time for phenological observations by information interpolation using regional amounts. Chapter 4 explains experimentally obtained results for our proposed method and a discussion of the results. Chapter 5 summarizes the contributions and future work.

2 RELATED WORK

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 data.

Phithakkitnukoon et al. [5] analyzed the behavior of travelers such as departure place, destination, and traveling means on a personal level in detail based on massive mobile phone GPS location records. Mislove et al. [6] developed a system that infers a Twitter user's feelings from Twitter text and visualizes changes of emotion in space–time. After research to detect events such as earthquakes and typhoons, Sakaki et al. [7] proposed a method to estimate real-time events from Twitter tweets. Cheng et al. [8] estimated 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 information interpolation is a new field.

3 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 in Japan. Our proposal is portrayed in Fig. 1.

We describe the best-time estimation method of organisms by analysis using a moving average method applied to geotagged tweets that include organism names. Section 3.1 describes how to collect geotagged tweets to be analyzed, whereas 3.2 describes preprocessing for conducting analysis, and 3.3 describes the best-time estimation method. In our proposed method up to now, the number of geotagged tweets has been small. It was possible to estimate the best time in a prefecture unit or municipality, but we were unable to analyze fine grain size. Therefore, using the method with information interpolation proposed in this paper, it is possible to estimate the best time to visit sightseeing spots with finer granularity. Section 3.4 presents an explanation of the information interpolation method, whereas 3.5 presents the output of the estimation result.

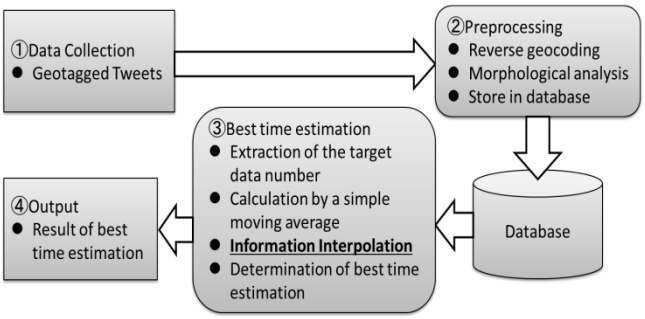

Figure 1: Our proposal summary.

3.1 Data Collection

This section presents a description of the Method of (1) data collection presented in Fig. 1. Geotagged tweets sent from Twitter are a 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 API [9] provided by Twitter Inc.

Next, we explain the number of collected data. According to a report presented by Hashimoto et al. [10], 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 used 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.

3.2 Preprocessing

This section presents a description of the method of (2) preprocessing presented in Fig. 1. Preprocessing includes reverse geocoding and morphological analysis, as well as database storage for data collected through the processing described in Section 3.1.

From latitude and longitude information in the individually collected tweets, reverse geocoding identified prefectures and municipalities by town name. We use a simple reverse geocoding service [11] that is 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 [12]. 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), and (。 / symbol)".

Preprocessing accomplishes the necessary data storage for the 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.

3.3 Estimating Best-Time Viewing

This section presents a description of the method of (3) best-time estimation presented in Fig. 1. Our method for estimating 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, estimated as the target word. The target word is a word including Chinese characters, hiragana, and katakana, which represents an organism name and seasonal change.

Next, we describe the simple moving average calculation, which uses a moving average of the standard of the best-time viewing judgment. It calculates a simple moving average on a daily basis using aggregate data by the target number of data extraction described above. Fig. 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.

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

$X(Y)$: 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 were a seven-day moving average and one-year moving average. A seven-day moving average is based on one week because tweets tend to be more numerous on weekends than on weekdays. In addition, 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 a simple moving average of the number of days specified for each organism to compare the seven-day moving average and a 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.

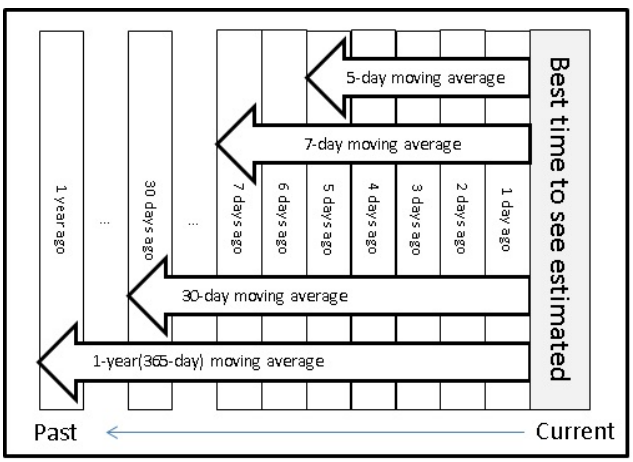

Figure 2: Number of days simple moving average.

As an example, we describe cherry blossoms. The Japan Meteorological Agency [13] 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" [14] is the first day on which blooming of 5–6 or more wheels of flowers occur on a specimen tree. The "Sakura in full bloom date" is the first day on which about 80% or more of the buds in the specimen tree are open. In addition, "Sakura" is the number of days from general flowering until full bloom: about five days. Therefore, "Sakura" in this study uses a five-day moving average as the standard.

Next, we describe an estimated judgment of the best time for viewing, as calculated using the simple moving average (seven-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 uses the number of tweets a day prior and a one-year moving average. Condition 1 is assumed to be satisfied when the number of tweets a day prior exceeds the one-year moving average, as shown in Formula 2.

Condition 2 uses a seven-day moving average and a biological moving average. The biological moving average varies depending on the organism that is estimated. It is five days in the case of cherry blossoms. For autumn leaves, it is 30 days. Therefore, in equation 3, we compare the seven-day moving average with the biological moving average, letting A be the long number of days, and letting B be the short number of days. In the case of estimation of cherry blossoms, A is 7 days; B is 5 days. For autumn leaves, A is 30 days; B is 7 days. Then we evaluate the moving average of A and B as shown in Equation 3. Furthermore, if the day on which Equation 3 holds lasts more than half of the number of days in A, Condition 2 is satisfied. In the case of cherry blossoms, A is 5 days. Therefore, condition 2 requires continuation for more than 3 days.

$$P_1 \geqq X(365) \quad (2)$$
$$X(A) \geqq X(B) \quad (3)$$

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

3.4 Information Interpolation Method

Herein, the information interpolation method will be described. Conventionally, we estimated the best time by application of the estimation method shown in the following estimated judgment using the moving average value described above. As a result, for analysis of a wide area such as a prefecture unit, the recall rate can be estimated as about 80%. However, with an estimate of granularity such as by sightseeing spots, an inability to estimate the viewing period because of a lack of data is a problem. Therefore, in this paper, we propose a method of using regional quantities that newly use information interpolation to compensate for the lack of data volume. The proposed method uses the result of reverse geocoding performed during preprocessing in the

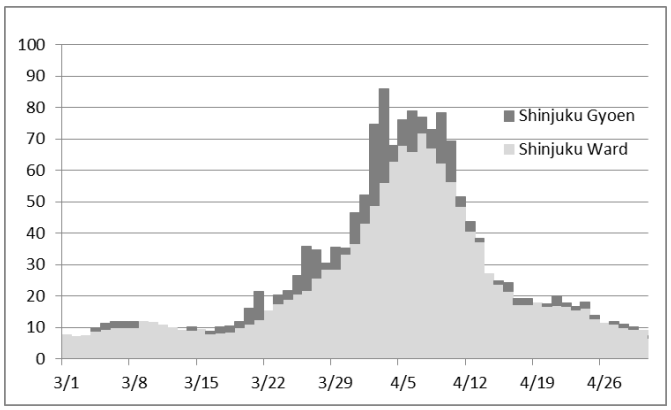

Figure 3: Example of information interpolation.

previous section. Tweets that were judged as the same municipality by reverse geocoding are totaled for each day by city, town, or village. Then, considering the characteristic by which the tweets move on a weekly basis, we obtain a seven-day moving average and set the seven-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.

As an example, we describe Shinjuku Gyoen, which is a cherry blossom sightseeing spot, and Shinjuku Ward, within which the spot is located. The dark gray area in Fig. 3 shows cherry blossom tweets related to Shinjuku Gyoen. An estimate might not be possible with just the number of tweets related to each sightseeing spot. For this reason, interpolation is performed using the seven-day moving average of tweets about cherry blossoms in Shinjuku Ward indicated by light gray in the city unit within which each sightseeing spot is located. In the proposed method, the best estimate is made using the number of tweets related to cherry blossoms at each tourist spot and the sum of the seven-day moving average of city unit.

However, if no tweet is related to sightseeing spots with the proposed method, estimation results of city unit are applied, so there are cases in which there is no difference depending on sightseeing spots in the same area. In the preliminary experiments, we succeeded in ascertaining the difference from nearby sightseeing spots if there are small tweets in the sightseeing spots.

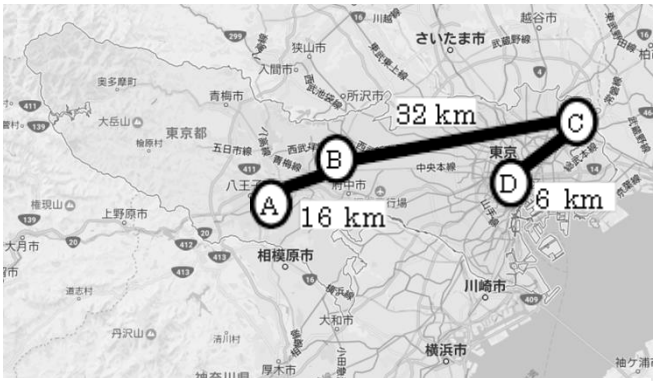

Figure 4: Position of target area.

3.5 Output

This section presents a description of the method of (4) output presented in Fig. 1. Output can be visualized using a best-time viewing result, as estimated by 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.

4 EXPERIMENTS

This chapter presents a description of the experiment to infer the best time to view cherry blossoms and autumn leaves for the proposed method described in Chapter 3. Section 4.1 describes the dataset used for optimal time reasoning. As an estimation result by sightseeing spot, section 4.2 presents the estimation result without using information interpolation, with the best estimation result obtained using information interpolation in section 4.3. Section 4.4 presents a comparison of the experimentally obtained results in Section 4.2 and Section 4.3.

4.1 Dataset

Datasets used for this experiment were collected using streaming API, as described for data collection in Section 3.1. Data are geotagged tweets from Japan during 2015/2/17 – 2017/8/31. The data include about 280 million items.

The estimation experiment to ascertain the best-time viewing of cherry blossoms uses the target word “cherry blossom,” which can be written as“桜” and“さくら”and “サクラ”in Japanese. For the experiment of autumn leaves, the target words are“紅葉,”“黄葉,”“コウヨウ,”“こうよう,”“モミジ,”and“もみじ”. We analyzed tweets that included a target word in the tweet text.

The following two experiments were conducted. The first is an experiment using the number of tweets including the target word and the sightseeing spot name without information interpolation. The second is an experiment using information interpolation. We use these datasets to estimate the optimum time for the sightseeing spots in Tokyo by experiments without information interpolation, (shown in Section 4.2) and experiments using information interpolation (shown in Section 4.3).

The subjects of the experiment were set as tourist spots in Tokyo. This report describes“Takao Mountain,”“Showa Memorial Park,”“Shinjuku Gyoen,”and“Rikugien.” Fig. 4 portrays the target area: 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.

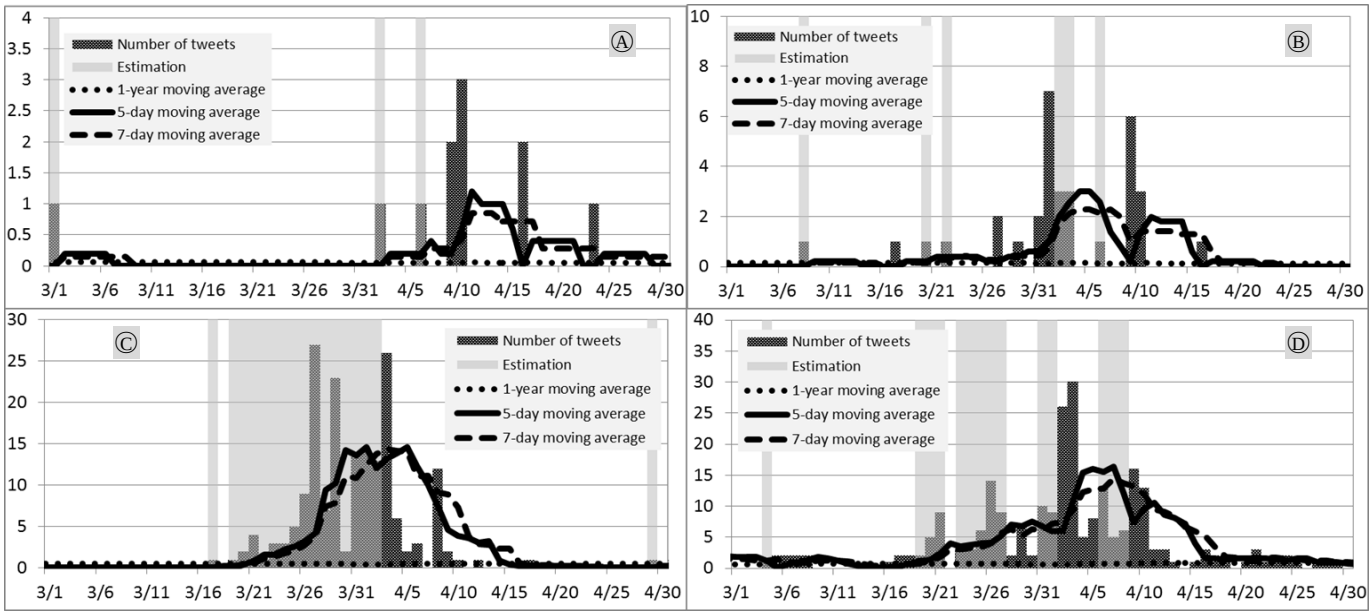

Figure 5: Experimental results obtained using tweets including the target word and the tourist spot name without interpolation (Cherry blossom).

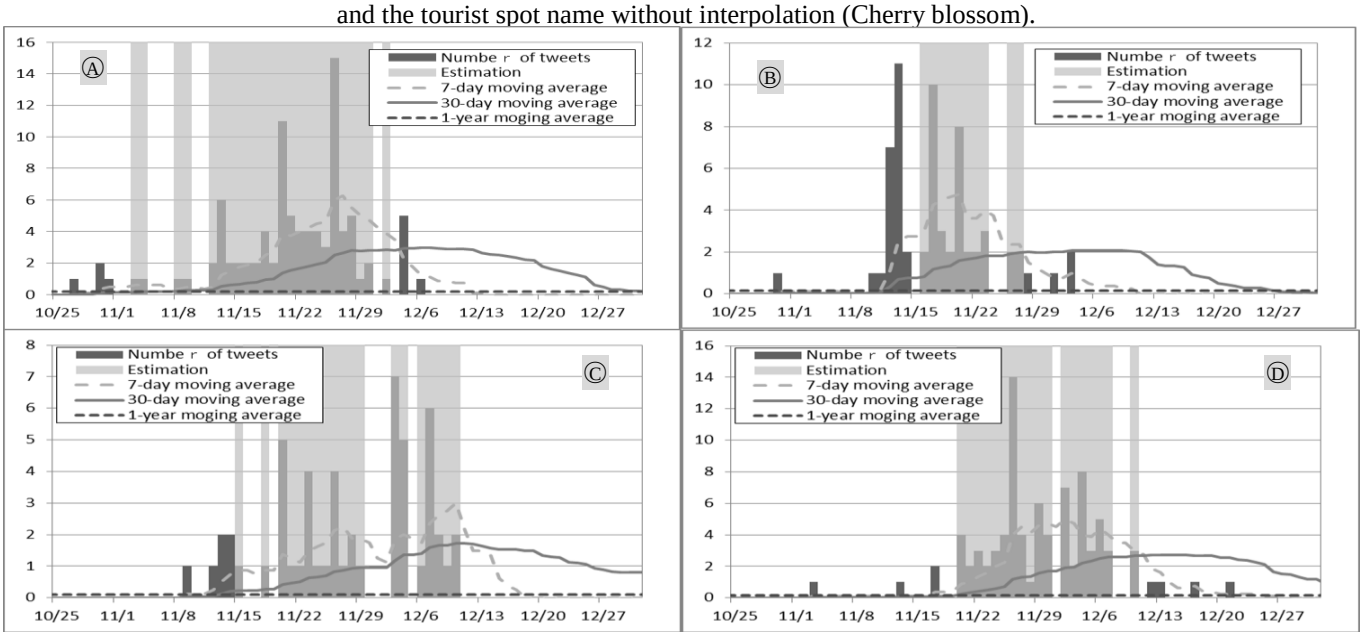

Figure 6: Experimental results obtained using tweets including the target word and the tourist spot name without interpolation (Autumn leaves)

4.2 Estimation Experiment for Best-Time Viewing without Information Interpolation

In this section, we present experimentally obtained results from estimating the best time without using information interpolation from tweets containing a target word and sightseeing spot name. Figure 5 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. In addition, the solid line shows a five-day moving average. The dashed line shows a seven-day moving average. The dotted line shows a one-year moving average.

At tourist spots targeted for the experiment in 2016, as portrayed in Fig. 5, 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 done using near-site estimation without interpolation. However, best-time viewing cannot be done in A and B because of the very small number of tweets.

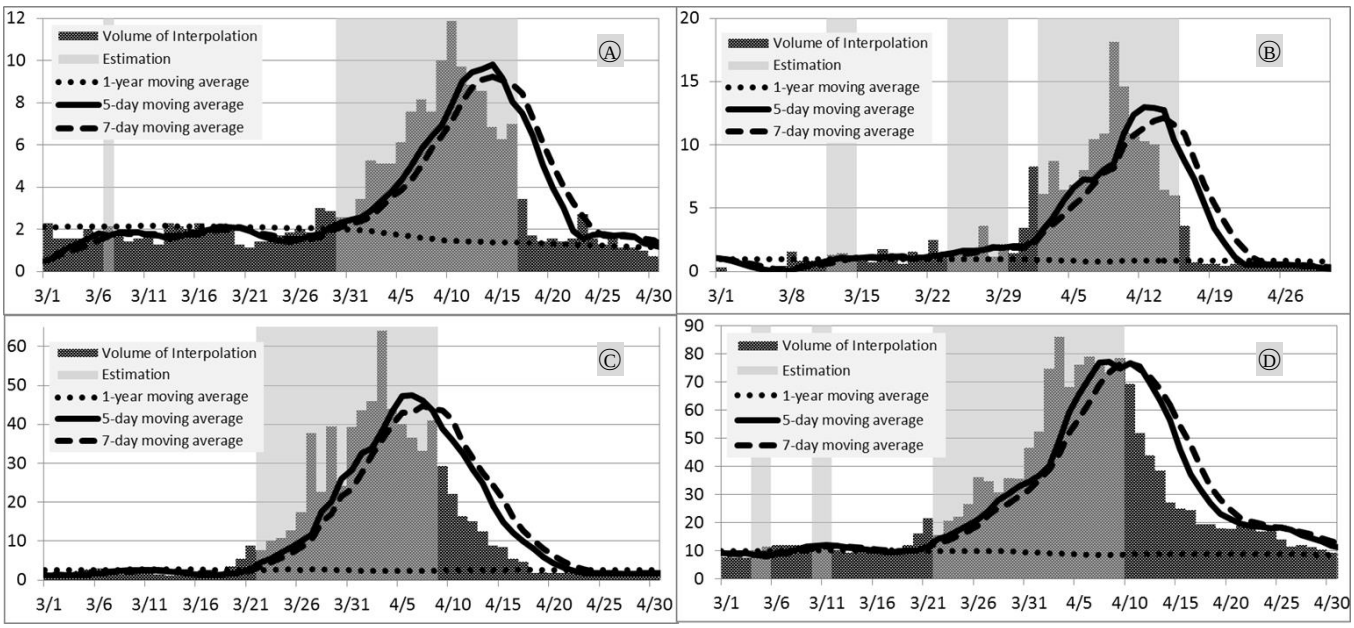

Figure 7: Experimental results obtained using tweets including the target word and the tourist spot name by interpolation

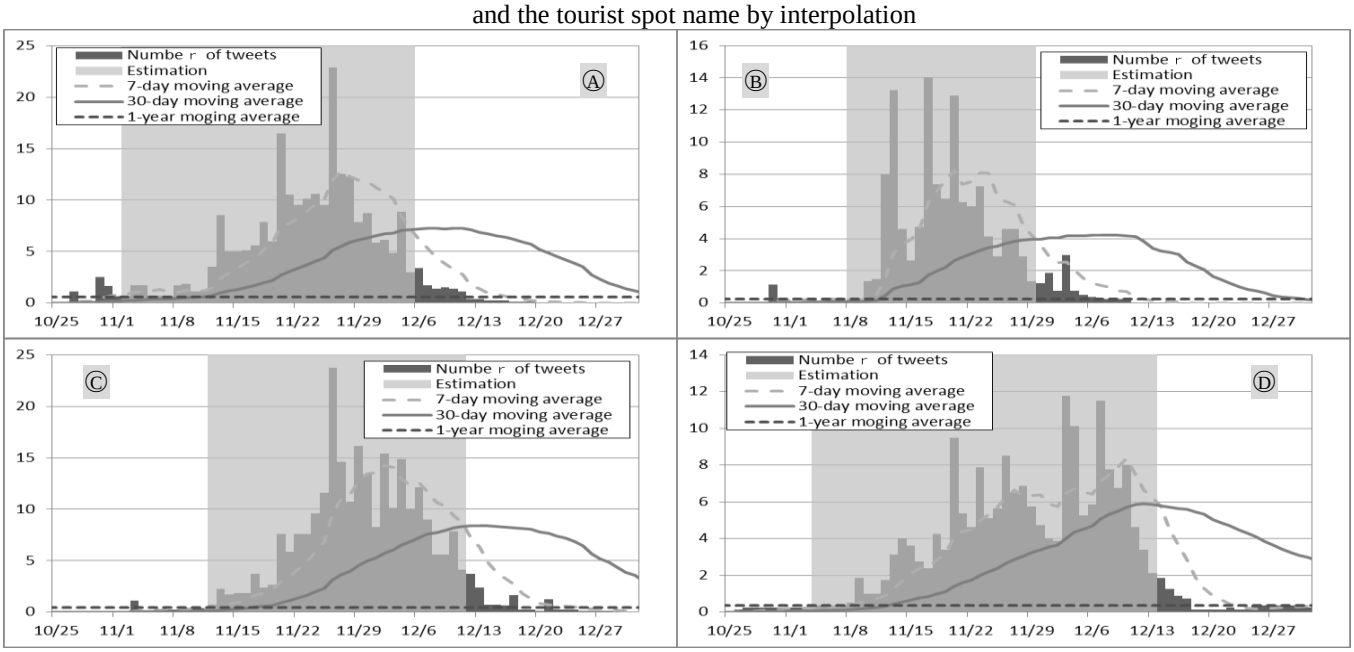

Figure 8: Experimental results obtained using tweets including the target word and the tourist spot name by interpolation

Next, we present experimentally obtained results of autumn leaves. Figure 6 portrays the estimated best time viewing results estimated using the word target word 'autumn leaves' in 2016. The notation in the figure is the same as that in Fig. 5. However, for cherry blossoms, the solid line is the five-day moving average, whereas the autumn leaves use a 30-day moving average. As shown in Fig. 6, the autumn leaves experiment has been estimated for each sightseeing spot because the viewing period is longer than that of cherry blossoms shown in Fig. 5. However, some parts cannot be estimated as continuous periods.

These results clarified that the method we proposed earlier cannot be predictive for detailed areas such as sightseeing spots. This result is attributable to the insufficient information volume.

4.3 Estimation Experiment for Best-Time Viewing with Information Interpolation

This section presents experimentally obtained results of estimation using information interpolation with regional quantities, which is the method proposed in this paper. Figure 7 presents results obtained using information interpolation for cherry blossom estimation. The notation is the same as the notation used in the previous section. Apparently, A and B can produce an estimate using the proposed method by increasing the number of tweets using information interpolation with surrounding tweets. In C and D, there are days that can be determined more accurately by interpolating the number of tweets.

Next, Fig. 8 presents results of information interpolation in autumn leaves estimation. Autumn leaves can be estimated as a continuous period using information interpolation. From this result, it can be inferred that the period estimation can be performed more accurately than without information interpolation.

These results demonstrate the possibility of resolving the difficulty of insufficient information when using sightseeing spot tweet data of the tourist spot area along with interpolation. One can then estimate the peak period for a particular tourist spot.

4.4 Comparing Best Time for Viewing Estimation and Observed Data

Table 1 presents a comparison of experimentally obtained results of estimating the best time using information interpolation. As the table shows, Experiment 1 used co-occurring words in tweets, including the sightseeing spot name coexisting with the target word "Sakura," without using the interpolation shown in 4.3. For Experiment 2, we used interpolation based on the information amount of the area including the tourist spots shown in 4.4. Numerical values in the table are the number of tweets including subject words and co-occurring words in Experiment 1. In Experiment 2, it is the sum of the number of tweets in Experiment 1 and the interpolation value of the regional information amount.

The light gray area in the table presents the date when the satiety prediction was made using the proposed method. In addition, confirming the flowering day and full bloom period of each sightseeing spot using JMA data is difficult.

Table 1: Comparison result in target areas of the best time to see the estimated and the observed data (Cherry blossom)

	Takao Mountain		Showa memorial park		Rikugien		Shinjuku gyoen	
	Exp.1	Exp.2	Exp.1	Exp.2	Exp.1	Exp.2	Exp.1	Exp.2
3/18	0	2.00	0	1.00	0	1.86	2	10.57
3/19	0	2.00	0	0.57	1	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	5	12.57	6	26.57
3/26	0	2.00	0	1.71	9	17.43	14	35.86
3/27	0	1.86	2	3.57	27	37.57	9	34.71
3/28	0	3.00	0	1.14	7	22.71	2	30.57
3/29	0	2.86	1	2.14	23	39.43	7	35.57
3/30	0	2.57	0	1.43	2	24.14	2	35.29
3/31	0	2.57	2	3.43	14	39.29	10	46.57
4/1	0	3.43	7	8.29	14	43.57	9	52.14
4/2	1	5.29	3	6.14	13	45.86	26	74.71
4/3	0	5.14	3	8.71	26	64.00	30	86.00
4/4	0	5.14	0	6.43	6	44.00	5	68.00
4/5	0	6.14	0	6.86	2	40.00	8	76.00
4/6	1	7.57	1	8.00	3	36.57	13	79.00
4/7	0	8.14	0	10.43	0	33.14	5	76.86
4/8	0	7.57	0	10.86	12	41.00	6	73.14
4/9	2	10.00	6	18.14	2	29.29	16	78.43
4/10	3	11.86	3	14.57	1	22.00	13	69.29
4/11	0	9.71	0	10.86	0	16.43	3	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	1	4.71	2	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	1	17.57
4/21	0	1.43	0	0.57	0	1.71	3	20.00
4/22	0	1.57	0	0.57	0	1.71	1	17.71
Precision	0.51	0.77	0.57	0.74	0.95	0.84	0.80	0.82
Recall	0.06	0.58	0.22	0.44	0.39	0.58	0.56	0.83

Nevertheless, this experiment to evaluate SNS data for flowering is valid also for weather forecasting companies [15] and for public service organizations [16] to evaluate optimum viewing times 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 for each target area for 2016 from 3/1 through 4/30 using formula (4) and formula (5).

$$Precision = \frac{Number\ of\ days\ to\ match\ the\ observed\ data}{Number\ of\ days\ in\ best\ time\ to\ see\ estimated} \quad (4)$$

$$Recall = \frac{Number\ of\ days\ to\ match\ the\ observed\ data}{Number\ of\ days\ of\ observation\ data} \quad (5)$$

We can explain the method using the example of Experiment 1 of Mt. Takao. The arrow portion of the flowering state is confirmed by hand as correct data, 1; the others are 0. In addition, the day estimated as the best time using the proposed method is set to 1; otherwise a day is 0. Furthermore, the percentage of days coinciding during 3/1 to

Table 2: Comparison result in target areas of the best time to see the estimated and the observed data (Autumn leaves)

	Takao mountain		Showa memorial park		Rikugien		Shinjuku gyoen	
	Exp.1	Exp.2	Exp.1	Exp.2	Exp.1	Exp.2	Exp.1	Exp.2
11/1	0	0.625	0	0.125	0	0	0	0.25
11/2	0	0.625	0	0.125	0	0	0	0.25
11/3	1	1.75	0	0.125	1	1.125	0	0.125
11/4	1	1.75	0	0.125	0	0.125	0	0.375
11/5	0	0.75	0	0.125	0	0.125	0	0.375
11/6	0	0.75	0	0.125	0	0.125	0	0.375
11/7	0	0.625	0	0.125	0	0.25	0	0.375
11/8	1	1.75	0	0.25	0	0.25	0	0.5
11/9	1	1.875	0	0.25	0	0.375	1	1.875
11/10	0	1.125	1	1.375	0	0.375	0	1
11/11	0	1.125	1	1.5	0	0.375	0	1
11/12	2	3.5	7	8	0	0.75	1	1.75
11/13	6	8.5	11	13.25	1	2.25	2	3.125
11/14	2	5	2	4.625	0	1.75	2	4
11/15	2	5	0	2.625	0	1.875	1	3.625
11/16	2	5.125	2	4.75	0	1.875	0	2.75
11/17	2	5.625	10	14	2	3.75	0	2.375
11/18	4	7.875	3	7.375	0	2.375	1	4.25
11/19	2	6	2	6.5	0	2.625	0	3.375
11/20	11	16.5	8	12.875	4	7.625	5	9.5
11/21	5	10.5	2	6.25	2	5.875	1	5.375
11/22	4	9.5	2	6	3	7.625	1	4.625
11/23	4	10.125	3	7.25	2	7.625	4	7.875
11/24	4	10.625	0	4.125	3	9.625	1	5.125
11/25	3	9.5	0	2.875	4	11.625	1	5.625
11/26	15	22.875	2	4.625	14	23.75	4	8.5
11/27	4	12.5	2	4.625	4	14.625	1	6.5
11/28	5	12.25	1	2.875	1	10.75	2	6.875
11/29	1	7.875	0	1.375	6	16.125	1	5.75
11/30	2	8.75	0	1.25	4	13.375	0	4.75
12/1	0	5.875	1	1.875	0	8.25	0	4
12/2	1	6.125	0	0.75	7	15.375	0	3.875
12/3	0	4.875	2	3	2	10.125	7	11.75
12/4	5	8.875	0	0.75	8	14.875	5	10.125
12/5	0	3	0	0.5	3	10	0	5.25
12/6	1	3.375	0	0.375	5	12.125	1	5.875
12/7	0	1.75	0	0.25	3	9	6	11.5
12/8	0	1.375	0	0.25	0	5.625	2	7.75
12/9	0	1.5	0	0.25	0	5.625	1	6.75
12/10	0	1.375	0	0.25	3	7.875	2	8
12/11	0	1.125	0	0	0	4.125	0	4.625
12/12	0	0.375	0	0	1	3.75	0	3.375
12/13	0	0.375	0	0	1	2.375	0	2.125
12/14	0	0.25	0	0	0	0.75	0	1.875
12/15	0	0.25	0	0	0	0.75	0	1.25
12/16	0	0.25	0	0	0	0.625	0	0.875
12/17	0	0.125	0	0	1	1.625	0	0.75
12/18	0	0.125	0	0	0	0.25	0	0.125
12/19	0	0.125	0	0	0	0.25	0	0.125
12/20	0	0	0	0	0	0.125	0	0.125
12/21	0	0	0	0	1	1.25	0	0.125
12/22	0	0	0	0	0	0.25	0	0.25
12/23	0	0	0	0	0	0.25	0	0.125
12/24	0	0	0	0	0	0.25	0	0.125
12/25	0	0	0	0	0	0.125	0	0.375
12/26	0	0	0	0	0	0.125	0	0.375
12/27	0	0	0	0	0	0.125	0	0.375
12/28	0	0	0	0	0	0.125	0	0.375
12/29	0	0	0	0	0	0	0	0.375
12/30	0	0	0	0	0	0	0	0.25
12/31	0	0	0	0	0	0	0	0.25
Precision	0.72	0.85	0.87	0.93	0.93	0.87	0.62	0.97
Recall	0.94	1.00	0.56	1.00	0.82	1.00	0.45	0.95

4/30 is shown. In Experiment 1 for Mt. Takao, except during the arrows, they match, but only 4/2 and 4/6 match during the arrow period. The Precision is 0.51 because the number of days matching the observed data is 31 days and the number of days of the best time for viewing is estimated is 61 days. In addition, because Recall is the ratio of matched days during the arrow, the number of days to match the observed data is 2 days and the number of days of observation data is 31 days. Therefore, it is 0.06.

Experimentally obtained results confirmed the tendency by which the relevance ratio and the recall rate became higher in Experiment 2 than in Experiment 1. In addition, A and B, which are at higher altitudes than C and D, exhibited 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 information interpolation along with regional data.

Table 2 presents a comparison of results of experiments for autumn leaves. The notation is the same as that of Table 1. The period was October 1 through December 31, 2016. The accuracy and recall ratio of experiment results obtained using information interpolation improved without information interpolation in each spot. Results confirm the effectiveness of the method proposed in this paper.

5 CONCLUSION

As described herein, to improve the best-time estimation accuracy and thereby enhance tourist information related to phenological observation, we proposed an information interpolation method. The proposed method showed the optimal time to view flowers at sightseeing spots by interpolating information using the seven-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, giving predictions in units required for sightseeing.

The results of cherry blossoms and autumn leaves experiments conducted for tourist spots in Tokyo in 2016 using the proposed method confirmed improvement of the estimation accuracy when using information interpolation. The proposed method using information interpolation for tweets related to target word might improve the real-world accuracy of estimating the best times. We confirmed the possibility of applying this proposed method to the estimation of viewpoints and lines of sight in areas and sightseeing spots with few tweets and little location information.

Although the proposed method showed success in interpolation of information and highly accurate estimation, it is necessary as a future task to verify whether the same result is obtainable also in biological seasonal observations other than those for cherry blossoms or autumn leaves. Future studies must also examine automatic extraction of target words and a method to perform future predictions in real time.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Japan Tourism Agency, “Tourism Nation Promotion Basic Law,” http://www.mlit.go.jp/kankocho/en/kankorikkoku/index.html (1, 2007).

[2] Ministry of Economy, Trade and Industry, “Study of landing type IT tourism and attract customers service,” http://www.meti.go.jp/report/downloadfiles/g70629a01j.pdf (3, 2007) (in Japanese).

[3] Twitter, “Twitter,” https://Twitter.com/ (4, 2014).

[4] M. Endo, Y. Shoji, M. Hirota, S. Ohno, and H. Ishikawa, “On best time estimation method for phenological observations using geotagged tweets,” IWIN2016 (2016).

[5] 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).

[6] A. Mislove, S. Lehmann, Y.Y. Shn, J.P. Onnela, and Rosenquist, “Understanding the Demographics of Twitter Users,” Proceeding. Fifth International AAAI Conference on Weblogs and Social Media (ICWSM’11), pp.133-140 (2011).

[7] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes Twitter users: real-time event detection by social sensors,” WWW 2010, pp.851-860 (2010).

[8] Z. Cheng, J. Caverlee, and K. Lee, “You are where you tweet: a content-based approach to geo-locating twitter users,” Proceedings of the 19th ACM International Conference on Information and Knowledge Management (2010).

[9] Twitter Developers, “Twitter Developer official site,” https://dev.twitter.com/ (4, 2014).

[10] 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).

[11] National Agriculture and Food Research Organization, “Simple reverse geocoding service,” http://www.finds.jp/wsdocs/rgeocode/index.html.ja (4, 2014).

[12] MeCab, “Yet Another Part-of-Speech and Morphological Analyzer,” http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html (9, 2012).

[13] Japan Meteorological Agency, “Disaster prevention information XML format information page,” http://xml.kishou.go.jp/ (12, 2011).

[14] Japan Meteorological Agency, “Observation of Sakura,” http://www.data.jma.go.jp/sakura/data/sakura2012.pdf (3, 2016).

[15] Weathernews Inc., “Sakura information,” http://weathernews.jp/sakura (3, 2016).

[16] Japan Travel and Tourism Association, “Whole country cherry trees,” http://sakura.nihon-kankou.or.jp (4, 2015).

(Received October 30, 2017)
(Revised April ,17 2018)

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.

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

Shigeyoshi Ohno earned M.Sci. and Dr. Sci. degrees from Kanazawa University and a Dr. Eng. degree from Tokyo Metropolitan University. He is currently a full Professor of Polytechnic University, Tokyo. His research interests include big data and web mining. He is a member of DBSJ, IPSJ, IEICE and JPS.

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.