Best-Time Estimation for Regions and Tourist Spots using Phenological Observations with Geotagged Tweets

Masaki Endo[*,**], Yoshiyuki Shoji[***], Masaharu Hirota[****], Shigeyoshi Ohno[*], and Hiroshi Ishikawa[**]

[*]Division of Core Manufacturing, Polytechnic University, Japan
{endou, ohno}@uitec.ac.jp
[**]Graduate School of System Design, Tokyo Metropolitan University, Japan
ishikawa-hiroshi@tmu.ac.jp
[***] Program-Specific Researcher, Academic Center for Computing and Media Studies, Kyoto University, Japan
shoji.yoshiyuki.5m@kyoto-u.ac.jp
[****] Department of Information Engineering, National Institute of Technology, Oita College, Japan
m-hirota@oita-ct.ac.jp

Abstract - In recent years, social network services (SNS) such as Twitter have become widely used, attracting great attention for many reasons. An important characteristic of Twitter is its real-time property. Twitter users post huge volumes of Twitter posts (*tweets*) related to daily events in real time. We assume that the tweet contents depend on the region, season, and time of day. Therefore, the possibility exists of obtaining valuable information for tourists from tweets posted during travel. As described in this paper, we propose a method to estimate regional best times for viewing flower blossoms from tweets including flower names. Our proposed method analyzes the number of tweets using a moving average. Additionally, we particularly examine geotagged tweets. Our experiments compare the best-time viewing estimated using our method to the flowering date and the full bloom date of cherry blossoms that the Japan Meteorological Agency has observed and posted. We conducted an experiment using data for the best-time viewing cherry blossoms during 2015 and 2016. Results confirmed that the proposed method can estimate the full bloom period accurately.

Keywords: trend estimation; phenological observation; Twitter

1 INTRODUCTION

In recent years, because of rapid performance improvement and the dissemination of various devices such as smart phones and tablets, diverse and vast data are generated on the web. Particularly, social networking services (SNS) have become prevalent because users can post data and various messages easily. According to the 2014 Communications Usage Trend Survey of the Ministry of Internal Affairs and Communications (MIC) [1], the percentage of Japanese people aged 13–39 years old using SNS is greater than 60%, the figure for people 40–49 years is higher than 50%. 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.

Here, we describe the provision of information to tourists using the web. Before SNS were used, local governments, tourism organizations, and travel companies provided regional tourism information using web pages. After SNS became widely used, they also undertook efforts to disseminate more detailed information related to respective tourist spots. The information is useful for tourists, but providing timely and topical travel information entails high costs for the information provider because they must update the information continually. Today, providing reliable information related to local travel is not only strongly demanded by tourists, but also local governments, tourism organizations, and travel companies, which bear high costs of providing the information.

Tourists also want real-time information and local unique seasonal information posted on web sites, according to a survey study of IT tourism and services to attract customers [3] by the Ministry of Economy, Trade and Industry (METI). Current web sites provide similar information in the form of guide books. Nevertheless, the information update frequency is low. Because local governments, tourism associations, and travel companies provide information about travel destination local unit independently, it is difficult for tourists to collect information for “now” tourist spots.

Therefore, providing current, useful, real-world information for travelers by capturing the change of information in accordance with the season and time zone of the tourism region is important for the travel industry. As described herein, we define "now" as information for tourism and disaster prevention required by travelers during travel, such as a best flower-viewing time and festivals and local heavy rains.

We propose a method to estimate the best time for phenological observations for tourism such as the best-time viewing cherry blossoms and autumn leaves in each region by particularly addressing phenological observations assumed for “now” in the real world. Tourist information for the best time requires a peak period, which means that the best time is not a period after and before falling flowers, but a period to view blooming flowers. Furthermore, the best times differ among regions and locations. Therefore, it is necessary to estimate a best time of phenological observation for each region and location. Estimating the best-time viewing requires the collection of much information having real-time properties. For this study, we use Twitter data obtained for many users throughout Japan.

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 estimation of best time of phenological observations. Chapter 4 describes experimentally obtained results for our proposed method and a discussion of the results. Chapter 5 summarizes the contributions and future work.

2 RELATED WORK

The amount of digital data is expected to increase greatly in the future because of the spread of SNS. Reports describing studies of the effective use of these large amounts of digital data are numerous. Some studies have used microblogs to conduct real-world understanding and prediction by analyzing information transmitted from microblogs. Kleinberg [5] detected a "burst" of keywords signaling a rapid increase in time-series data. Ochiai et al. [6] proposed a disambiguation method for family names that are also used as place names using dynamic characteristic words of topics that vary from period to period, including static characteristic words and locations that are independent of specific seasonal variation according to the location as a target of microblog. Kurata et al. [7] developed a system to detect events in real space using geotagged tweets. This system can grasp what events occur in time and place by the top 10 of frequent word extraction conducted in each time zone. Sakaki et al. [8] proposed a method to detect events such as earthquakes and typhoons based on a study estimating real-time events from Twitter. Kunneman et al. [9] proposed a method to differentiate among tweets posted before, during, and after a soccer match using machine learning. Hurriyetoglu et al. [10] proposed a method to estimate the time to a future soccer match using tweet streams with local regression over a word time series. Tops et al. [11] described a method to classify the time to an event in automatically discretized categories using support vector machines. Consequently, various methods for extracting event and location information have been discussed. Nevertheless, a method used to estimate the start and end of the full bloom period of phenological observations using tweets is controversial.

3 OUR PROPOSED METHOD

This chapter presents a description of a method of analysis for target data collection and our best-time estimation to get a guide for phenological change from Twitter in Japan. Our proposal is portrayed in Fig. 1.

3.1 Data Collection

This section presents a description of the Method of (1) data collection shown in Fig. 1. Geotagged tweets sent from

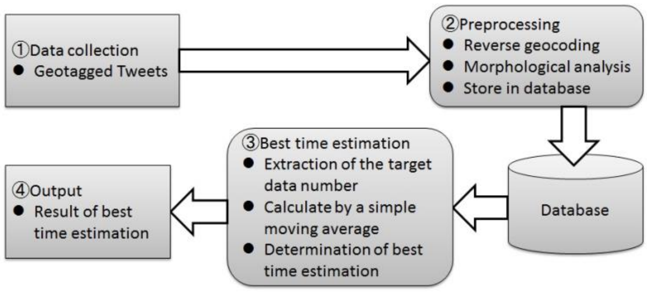

Figure 1: Our proposal summary

Table 1: Transition example 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

Twitter are a collection target. Range geo-tagged tweets include the Japanese archipelago (120.0 ≤ longitude ≤ 154.0 and 20.0 ≤ latitude ≤ 47.0) as the collection target. Collection of these data was done using Streaming API [12], one API provided by Twitter, Inc.

Next, we describe the collected number of data. The percentage of geotagged tweets among tweets originated in Japan, according to the study of Hashimoto et al. [13] as a whole is about 0.18%. Such tweets are very few among all data. However, the collected geo-tagged tweets, shown as an example in Table 1, number about 70,000, even on weekdays. On weekends there are also days on which more than 100,000 such messages are posted. We use about 30 million geo-tagged tweets from 2015/2/17 through 2016/6/30. For each day of collection, the number during the period covered was about 72,000. We calculated the best time for flower viewing, as estimated by processing the following sections using these data.

3.2 Preprocessing

This section presents a description of the method of (2) preprocessing shown 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.

Reverse geocoding identified prefectures and municipalities by town name from latitude and longitude information of the individually collected tweets. We use a simple reverse geocoding service [14] available from the National Agriculture and Food Research Organization in this process: e.g., (latitude, longitude) = (35.7384446, 139.460910) 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 [15]. 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 performs the necessary data storage for the best-time viewing, as estimated Based on results of the processing of the data collection and 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 the Best-Time Viewing

This section presents a description of the method of (3) best-time estimation presented in Fig. 1. Our method for estimating the best-time viewing processes the target number of extracted data and calculates a simple moving average, yielding an inference of the best flower-viewing time. The method defines a word related to the best-time viewing, estimated as the target word. The target word can include Chinese characters, hiragana, and katakana, which represents an organism name and seasonal change, as shown in Table 2.

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

Table 2: Examples of the target word

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

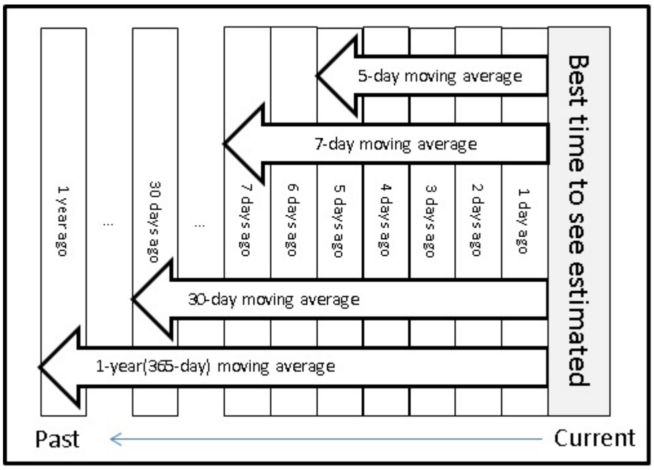

Figure 2: Number of days simple moving average

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 $d$ay 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 7-day moving average and 1-year moving average. As shown in Table 1, since geo tag tweets tend to be more frequent at weekends than on weekdays, a moving average of 7 days is taken as one of estimation criteria. And the phenomenological observation is based on the one-year moving average as the estimation criterion because there are many "viewing" events every year, such as "viewing of cherry blossoms", "viewing of autumn leaves" and "harvesting month".

In addition to the 7-day moving average and the 1-year moving average, we also explain the moving average of the number of days specified for each phenological. In this study, we set the number of days of moving average from specified biological period of phenological.

As an example, we describe cherry blossoms. The Japan Meteorological Agency (JMA) [16] carries out phenological observations of "Sakura," which yields two output items of the flowering date and the full bloom date observation target. "Sakura of flowering date" [17] is the first day of blooming 5–6 or more wheels of flowers of a specimen tree. "Sakura in full bloom date" is the first day of a state in 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 5 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 viewing, which was calculated using the simple moving average (7-day moving average, 1-year moving average, and another biological moving average). It specifies the two conditions as a condition of an estimated decision for the best-time viewing.

Condition 1 uses the 1-year moving average and the numbe r of tweets containing the organism name of each day. Comp are the number of tweets on each day and the 1-year moving average calculated for each day as shown in Equation 2. Th e day when the number of tweets on each day exceeds the 1-year moving average is the day when the condition1 holds.

$$P_1 \geqq X(365) \quad (2)$$

For condition 2, we use equation 3 to make a judgment using 7-day moving average and biological moving average. Here, A in Equation 3 refers to the shorter number of days by comparing "moving average of 7 days" and "moving average of bioequivalence". B is a long number of days.

$$X(A) \geqq X(B) \quad (3)$$

As an example, cherry blossoms use a 7-day moving average and 5-day moving average, so A is 5 days and B is 7 days. This determines the date on which the moving average of a short number of days exceeds the moving average of a long number of days. Subsequently, it is assumed that Condition 2 is satisfied when the moving average of a short number of days exceeds the moving average of a long number of days continuously. The number of consecutive days was made equal to or more than half of the moving average of a short number of days. In the case of cherry blossoms, the 5-day moving average is shorter than the 7-day moving average. Therefore, 5 days / 2 = 2.5 days ≒ 3 days as a standard. If the 5-day moving average exceeds the 7-day moving average by 3 days or more, it shall be the date satisfying Condition 2.

Finally, we estimate the day that both Condition 1 and Condition 2 are satisfied as best time to see.

3.4 Output

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

4 EXPERIMENTS

In this chapter, the experimental explanation for guessing the optimum time to see the flower in the proposed method described in Chapter 3 is shown. This shows the dataset used for the optimum time reasoning to see flowers that bloom completely in section 4.1. Section 4.2 shows the target word and target area used in the experiment. Section 4.3 shows experimental results on 2015 cherry blossoms viewing and 2016 cherry blossoms viewing. Section 4.4 shows the result of comparing the estimation result with observation data. Section 4.5 shows the experiment results using the sightseeing spots co-occurring in the target word.

4.1 Dataset

Datasets used for this experiment were collected using streaming API, as described for data collection in Section 3.1. Data are geo-tagged tweets from Japan during 2015/2/17 – 2016/6/30. The data include about 30 million items. We are using these datasets for experiments to infer the best time for cherry blossom viewing in 2015 and in 2016.

4.2 Estimation Experiment for Best-Time Viewing Cherry Blossoms

Figure 3: Position of target area

The estimation experiment to ascertain the best-time viewing cherry blossoms uses the target word in Table 2: "Sakura". The target word is “cherry blossom,” which is “桜” and “さくら” and “サクラ” in Japanese. The experimental target areas were "Tokyo," "Ishikawa", "Kyoto," and “Hokkaido." For each area, a specimen tree is used for observations by the JMA. The cities of “Chiyoda," "Kanazawa," "Kyoto," and "Sapporo" are target areas. In addition, an experiment using co-occurrence words was conducted using tweets with the item "Sakura" for many tourist spot named in Table 2, which also shows with the number of occurrences in respective regions. As described in this paper, we specifically examined "Shinjuku Gyoen", "Rikugien Garden," "Goryokaku," and "Kenroku-en".

Figure 3 presents the target area location. Kyoto and Hokkaido are separated by about 1,000 km straight line distance. Kyoto and Tokyo are about 360 km apart. Because of their latitudes, cherry trees flower later in the north in Hokkaido than in Kyoto. Moreover, higher altitudes and consequently cooler temperatures delay flowering even when locations have similar latitudes. Although issues related to altitude were not particularly addressed in this study, they are not expected to affect important results for single sites.

4.3 Target Word Results in Target Areas

Figure 4 presents experimentally obtained results for the estimated best-time viewing in 2015 using the target word cherry blossoms in the target area of "Tokyo." The dark gray bar in the figure represents the number of tweets. The light gray part represents the best-time viewing as determined using the proposed method. In addition, the solid line shows a 5-day moving average. In addition, the solid line shows a 5-day moving average. The dashed line shows a 7-day moving average. The dotted line shows the 1-year moving average. Figure 5 shows the estimated best-time viewing, as inferred from experimentally obtained results in 2016 using the target word cherry blossoms in the target area of "Tokyo."

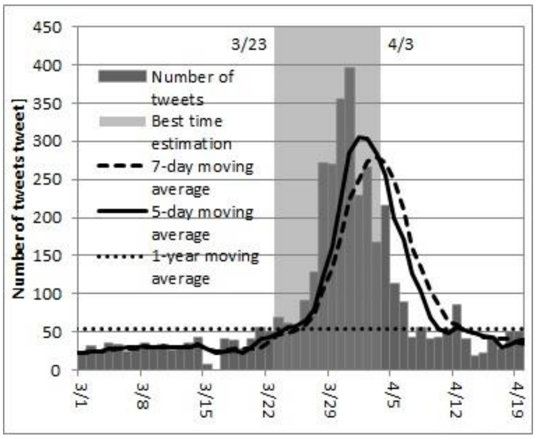

Figure 4: Results of the best time to see, as estimated by Tokyo (2015)

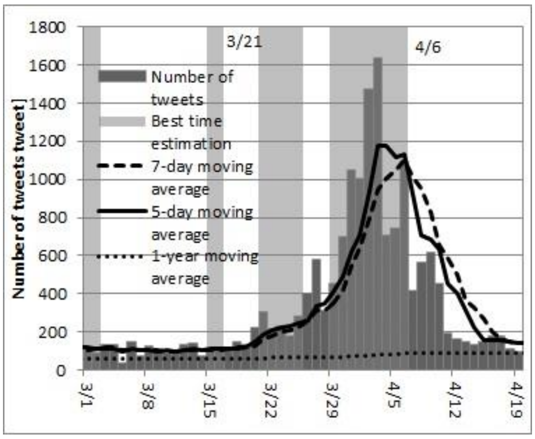

Figure 5: Results of the best time to see, as estimated by Tokyo (2016)

For Tokyo in 2015, as portrayed in Fig. 4, we obtained the greatest number of data. The greatest number of tweets per day reached about 400. Our proposed method indicates the best-time viewing as 3/23 – 4/3. Condition 1 shown in 3.3 is the day when a dark gray bar exceeds the dotted line. Condition 2 is the day when the solid line exceeds the broken line for more than 3 days. Therefore, in our proposed method, we estimated 3/23 – 4/3 which satisfy both condition 1 and condition 2 as best-time viewing.

1Our proposed method shows the best-time viewing as 3/21 – 4/6 in Fig. 5. The estimation for the best-time viewing in 2016 indicates a longer period than that in 2015, which is consistent with the trend of 2016, with low-temperature days after flowering. Tokyo of 2016, as presented in Fig. 5, also has the largest number of data in the area of the experimental subjects of 2016. More than 1,600 tweets were sent on some days, which is about four times that of 2015. Therefore, the 1-year moving average value for the rapid increase in the number of tweets is reduced. For that reason, much noise is included in the estimate of the best-time viewing.

Figure 6 presents results of 2015 for Ishikawa Prefecture. Results of 2016 for Ishikawa Prefecture are portrayed in Fig. 7. The greatest numbers of data were, respectively, 15 tweets and 45 tweets. Ishikawa data are far fewer than those of Tokyo. However, 2015 has been the best-time viewing

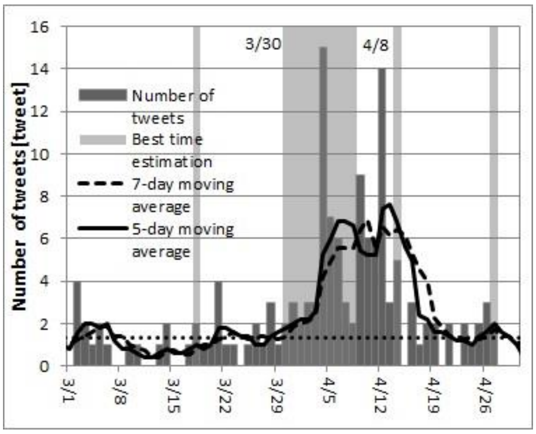

Figure 6: Results of best time to see, as estimated by Ishikawa (2015)

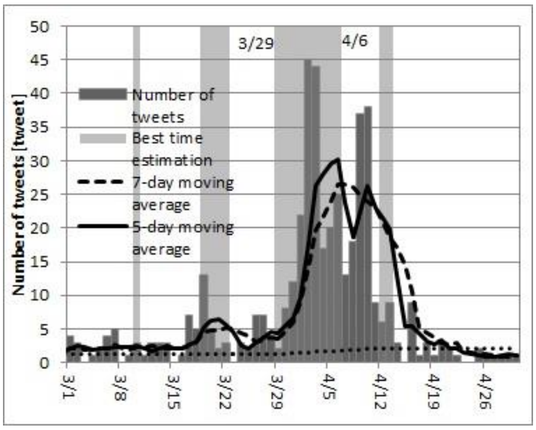

Figure 7: Results of best time to see, as estimated by Ishikawa (2016)

was estimated as 3/30 – 4/8. In 2016, the best-time viewing was estimated as 3/29 – 4/6. Noise contents that are unrelated to "Sakura" organisms are also included in tweets. However, before the peak period, tweets abound for budding and flowering cherry. After the peak period, tweets related to cherry blossom leaves are prominent.

From the above, it is possible to estimate the best time by using the proposed method only using the number of occurrences of the target word. This is because the cherry trees of the target word shown in this experiment are the most frequent tweets pointing to cherry trees of creatures in Japan. In the experiment, the accuracy in the area with many tweets was relatively high. However, in addition to tweets as cherry trees of living things, it is considered that there are also tweets including cherry blossoms that become noise used for different purposes such as person's name and food. Also, in the case of another target word, there are cases where the best time to estimate the best time other than the creature is erroneously estimated. Although this paper does not mention analysis of tweet contents, it is a future subject. Also, in the case of other target words, there are cases where the best time other than the organism is erroneously estimated. Although this paper does not mention analysis of the contents of tweets, it is a future task.

4.4 Comparing Best Times for Viewing Estimation and Observed Data

Table 3 presents results of comparison between the estimated best-time viewing and JMA observation data in target area in 2015. Dates in the table are the target dates for the estimated best-time viewing. The thin gray portion of each region is the day determined as the best-time viewing: as an example, Tokyo's best-time viewing in 2015 was 3/23 – 4/3. This result represents the same day and the estimated best-time viewing thin gray part of the previous section in Fig. 4. Furthermore, the arrow indicates the period of up to "cherry blossoms in full bloom date" from "cherry flowering date" that the JMA has observed in each region. As an example, Tokyo observations are based on the specimen tree in Chiyoda. The "Sakura flowering date" of 2015 is 3/23. The "Sakura in full bloom date" of 2015 is 3/29. Specimen trees of the JMA of the experimental target area are the following. Ishikawa is Kanazawa. Kyoto is Kyoto City. Hokkaido is Sapporo. Recall and precision using the observed data and the estimated best-time viewing results are calculated for each target area for 2015 from 3/1 – 6/30 using formula (4) and formula (5).

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

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

The precision ratio average in Table 3 is about 20%. A low precision ratio does not include the period from full bloom to abscission. The best-time viewing is estimated as 3/30 – 4/3 for Tokyo as determined by the JMA as the best-time viewing after full blooming of cherry trees. Therefore, the result presents the possibility of providing the best-time viewing information that is necessary to complement tourist observation data of the JMA using the proposed method. However, the data are few for areas such as Kanazawa. Therefore, the moving average used to estimate the best time for flower viewing is vulnerable to extreme changes.

Moreover, Hokkaido, Tokyo, Ishikawa, and Kyoto recall is higher than that of municipal districts. These experiments use aggregate data of each whole area against observation data of a sample tree of the JMA. Chiyoda and Kanazawa are regions within prefectures. They therefore have a low recall rate because the data are fewer. Kyoto and Sapporo show no decrease of recall because many data in the region are city data. Results of this best-time estimation should be provided as tourist information in each region for which there is limited information of target areas.

Table 4 presents experimentally obtained results for 2016. The notation is the same as that used in Table 3. The experimental period in 2016 was 3/1 – 6/30. Data of 2016 obtained using our proposed method were also confirmed best-time estimation for each region. Data confirmed the best-time estimation after full bloom observation by the JMA. Compared to 2015, 2016 was confirmed to have a long best-time duration because of low temperatures after flowering. However, precision and recall for some data loss are lower than in 2015.

From the above, the proposed method is useful for estimating the optimum time for viewing cherry blossoms in areas where about 10 tweets per day were obtained. However, since the area under study in this experiment is the capital of the prefecture, there are also relatively many tweet data. There is also a specimen tree used for JMA observation. However, in many other areas there may be regions where there are few data. Therefore, we need to verify further in other areas.

4.5 Results of the Full Bloom Estimation using the Co-occurrence Word

Table 5 presents estimation results for the best-time viewing in 2015 with tweets that include the tourist spot name co-occurring with the target word "Sakura". The co-occurrence words are tourist spot names used for estimation with the proposed method: "Shinjuku Gyoen," "Rikugien Garden," "Goryokaku," and "Kenroku-en". The numerical values in the table are the numbers of tweets including the target word and the co-occurrence word. The light gray part shows a date for which full bloom estimation was made

Table 3: Comparison result in target areas of the best time to see the estimated and the observed data (2015)

日付	Tokyo	Chiyoda	Ishikawa	Kanazawa	Kyoto	Kyoto city	Hokkaido	Sapporo
3/18								
3/19								
3/20								
3/21								
3/22								
3/23								
3/24								
3/25								
3/26								
3/27								
3/28								
3/29								
3/30								
3/31								
4/1								
4/2								
4/3								
4/4								
4/5								
4/6								
4/7								
4/8								
4/21								
4/22								
4/23								
4/24								
4/25								
4/26								
Precision	33.3%	22.2%	35.7%	5.9%	7.1%	6.3%	23.8%	27.8%
Recall	100.0%	57.1%	100.0%	20.0%	25.0%	25.0%	100.0%	100.0%

Table 4: Comparison result in target areas of the best time to see the estimated and the observed data (2016)

日付	Tokyo	Chiyoda	Ishikawa	Kanazawa	Kyoto	Kyoto city	Hokkaido	Sapporo
3/18								
3/19								
3/20								
3/21								
3/22								
3/23								
3/24								
3/25								
3/26								
3/27								
3/28								
3/29								
3/30								
3/31								
4/1								
4/2								
4/3								
4/4								
4/5								
4/6								
4/7								
4/8								
4/24								
4/25								
4/26								
4/27								
4/28								
4/29								
4/30								
5/1								
Precision	44.4%	53.3%	40.0%	46.2%	52.4%	44.0%	40.0%	44.4%
Recall	72.7%	72.7%	100.0%	100.0%	100.0%	100.0%	66.7%	66.7%

Table 5: Best time to see the estimated and tweet the number of tourist spot name to co-occurrence (2015)

Date	Shinjuku Gyoen	Rikugien Garden	Goryokaku	Kenroku-en
3/17	0	0	0	0
3/18	1	0	1	0
3/19	0	0	0	0
3/20	0	0	0	0
3/21	1	0	0	0
3/22	0	0	0	0
3/23	0	0	0	0
3/24	3	0	0	0
3/25	0	0	0	0
3/26	0	0	0	0
3/27	0	4	0	0
3/28	0	4	0	0
3/29	3	2	0	0
3/30	5	2	0	0
3/31	1	3	0	0
4/1	4	1	0	0
4/2	2	1	0	0
4/3	0	1	0	0
4/4	2	0	0	2
4/5	1	0	0	2
4/6	0	0	0	0
4/7	0	0	0	0
4/8	0	0	0	0
4/9	0	0	0	1
4/10	0	0	0	0
4/11	0	0	0	1
4/12	1	0	0	2
4/13	0	0	0	0
4/14	0	0	0	1
4/15	0	0	1	0
4/16	0	0	0	0
4/17	2	0	1	0
4/18	4	0	0	0
4/19	1	0	0	1
4/20	0	0	1	0
4/21	1	0	1	0
4/22	0	0	0	0
4/23	0	0	0	0
4/24	0	0	1	0
4/25	0	0	2	1
4/26	0	0	3	0
4/27	0	0	1	0
4/28	0	0	0	0
4/29	0	0	0	0
4/30	0	0	0	0
Precision	0.0%	0.0%	100.0%	100.0%
Recall	0.0%	0.0%	14.3%	33.3%

Table 6: Best time to see the estimated and tweet the number of tourist spot name to co-occurrence (2016)

Date	Shinjuku Gyoen	Rikugien Garden	Goryokaku	Kenroku-en
3/17	1	1	0	0
3/18	2	0	0	0
3/19	2	0	0	0
3/20	5	2	0	0
3/21	9	4	0	0
3/22	0	1	0	0
3/23	3	3	0	0
3/24	3	3	0	0
3/25	5	4	0	0
3/26	12	9	0	0
3/27	7	26	0	0
3/28	1	7	1	1
3/29	5	18	4	0
3/30	9	18	0	1
3/31	6	14	0	1
4/1	6	13	1	2
4/2	22	13	1	7
4/3	29	21	0	7
4/4	4	5	0	6
4/5	6	2	0	3
4/6	9	3	0	6
4/7	5	0	0	3
4/8	5	13	0	2
4/9	12	2	0	3
4/10	13	1	0	7
4/11	2	0	0	1
4/12	3	1	0	0
4/13	1	0	0	1
4/14	0	0	0	0
4/24	0	0	0	0
4/25	1	0	2	0
4/26	0	0	3	0
4/27	0	0	3	0
4/28	1	0	3	0
4/29	1	1	5	0
4/30	0	0	6	0
5/1	0	1	4	0
5/2	2	0	9	0
5/3	1	0	8	0
5/4	0	0	5	0
5/5	0	0	2	0
5/6	0	0	6	0
5/7	1	0	3	0
5/8	0	0	0	0
5/9	0	0	2	0
5/10	1	0	0	0
Precision	50.0%	100.0%	66.7%	66.7%
Recall	22.2%	27.8%	23.5%	12.5%

using the proposed method. Confirmation of the flowering date and the full bloom date of each tourist spot is difficult using JMA data. Therefore, verification of this experiment was used to assess flowering and the best-time viewing the sights according to services or blogs, in addition to SNS of weather information companies [18] and public interest institutes [19]. The arrow representing the time to bloom from flowering was confirmed manually at each tourist spot.

Table 5 shows that the data of each tourist spot is very few, but one can confirm the differences of full bloom times for tourist spots. Even in the vicinity of each other like "Shinjuku Gyoen" "Rokugien", time difference can be seen. This result is different from estimation by JMA which depends on observation of specimen tree. The proposed method shows the possibility to estimate the best-time viewing date of each tourist spot. However, in the proposed method, the accuracy of extracting best-time viewing as a period is low. For that reason, it is a big task in the future as to estimation in regions and tourist spots with few tweets.

Table 6 presents experimentally obtained results for 2016. The notation is the same as that used in Table 5. Similar to Tables 4 and 5 presented in earlier sections, 2016 had a longer best-time duration than that in 2015 because of lower temperatures after flowering. Table 5 shows that the data are increasing for each tourist spot. Therefore, the estimated best-time results obtained for 2016 using the proposed method tend to match the best-time viewing in 2015 indicated by the arrow. However, the best time estimation of the tourist resort including the optimum viewing time estimated using the proposed method increased the amount of data compared to 2015, so it was possible to confirm the improvement in accuracy, but further improvement of the method is necessary.

5 CONCLUSION

This paper suggested using Twitter to generate a useful approach to estimate the best time to present sightseeing information related to phenology observation. In the proposed method, we used a geo-tagged tweet containing the organism name of the target word to infer the optimal time to see flowers in Japan. The result of the cherry blossom experiment shows that the seasonal change of the tweet and the actual seasonal change are related to the estimate. Therefore, the proposed method presents the possibility to estimate the best time in the real world by observing the tweet related to the organism name. By using this, we are considering application to a system that can judge whether the phenomenon will become a tourist target when visiting sightseeing by checking whether the phenology is the best state in the area. Also, the granularity of the proposed method differs depending on the target word, region, and sightseeing spot. In this paper, we conducted experiments on sightseeing spots that co-occur with prefectures and target words. The results confirmed the possibility of displaying tourist information in real time for each area and sightseeing spot by estimating the optimum time using geotagged tweets. On the other hand, further consideration is needed on the estimation of the best time in areas and sightseeing spots where the number of data is small. Future research should verify that the proposed method is applicable to other organisms. Depending on the target word, there are many false positive cases, so in the future we will also consider methods of analyzing tweets contents. By extending the proposed method, we would like to connect to a system that allows travelers to obtain event information and disaster information on travel destinations in real time.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] Ministry of Internal Affairs and Communications, "2014 Communications Usage Trend Survey Results," http://www.soumu.go.jp/johotsusintokei/statistics/data/150717_1.pdf (7, 2015) (in Japanese).

[2] Twitter, "Twitter," https://Twitter.com/ (4, 2014).

[3] 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).

[4] Manabu Okumura, "Microblog Mining," IEICE Technical Report NLC2011 −59 (2012) (in Japanese).

[5] 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).

[6] K. Ochiai, and D. Torii, "Toponym Disambiguation Method for Microblogs Using Time-varying Location-related Words," IPSJ TOD, Vol.7, No.2, pp.51-60 (2014) (in Japanese).

[7] A. Kurata, K. Uehara, and J. Murai, "Auto situation detecting system using Twitter," The 75th National Convention of JPSJ, 1V-1, pp.97-98 (2013) (in Japanese).

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

[9] Florian Kunneman and Antal van den Bosch., "Leveraging unscheduled event prediction through mining scheduled event tweets," BNAIC2012, p.147 (2012).

[10] Ali Hurriyetoglu, Florian Kunneman, and Antal van den Bosch, "Estimating the time between twitter messages and future events," DIR-2013, pp.20–23 (2013).

[11] Hannah Tops, Antal van den Bosch, and Florian Kunneman, "Predicting time-to-event from Twitter messages," BNAIC 2013, pp.207-2014 (2013).

[12] Twitter Developers, "Twitter Developer official site, " https://dev.twitter.com/ (4, 2014).

[13] Y. Hashimoto, 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).

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

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

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

[17] Japan Meteorological Agency, "observation of Sakura," http://www.data.jma.go.jp/sakura/data/sakura2012.pdf (3, 2015).

[18] Weathernews Inc., "Sakura information," http://weathernews.jp/koyo (9, 2015).

[19] Japan Travel and Tourism Association, "whole country cherry trees," http://sakura.nihon-kankou.or.jp/ (4, 2015).

(Received October 8, 2016)
(Revised February 16, 2017)

Masaki Endo

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 Assistant 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.

Yoshiyuki Shoji
Yoshiyuki Shoji is a program-specific Researcher at Kyoto University, Japan. He earned a Ph.D. Degree in Informatics from Kyoto University in 2015. He worked as a researcher at Kyoto University in 2015, and worked as a program-specific Assistant Professor at Tokyo Metropolitan University in 2016. His research interests include social informatics, information retrieval, and web mining.

Masaharu Hirota
Masaharu Hirota earned his Doctor of Informatics degree in 2014 from Shizuoka University. Since April 2015, he has been working as an Assistant Professor at the National Institute of Technology, Oita College, Department of Information Engineering. His main research interests include geo social data, multimedia, and visualization. He is a member of ACM, IPSJ, and IEICE.

Shigeyoshi Ohno
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
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.