Co-occurring Word Determination Used for Estimating Best Times for Viewing Cherry Blossoms

Yusuke Takamori
Electronic Information Course
Polytechnic University
Kodaira-shi, Tokyo
e-mail: b19309@uitec.ac.jp

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

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

***Abstract*—As described herein, we propose a method to increase the amount of data used to estimate the best time for viewing seasonal organisms, particularly flowers. Observations of the best viewing times of seasonal organisms, conducted by the Japan Meteorological Agency, are required by those in the tourism industry and tourists, but the numbers of such official observations are decreasing: an accurate alternative is needed. As one alternative, we have investigated estimation of the best viewing seasons and times of biological organisms using Twitter, which is widely used in Japan. Specifically addressing difficulties posed by the decrease in data for estimating the best viewing times of biological organisms, which was a difficulty of earlier research, we propose a new method for improvement. The proposed method using co-occurring words is used for obtaining tweets related to seasonality. Combining the proposed method with the conventional method for estimating the best viewing times of seasonal organisms showed improved accuracy compared to that achieved using the conventional method alone.**

Keywords- SNS; Twitter; estimating; cherry blossoms; co-occurrence.

I. INTRODUCTION

Along with the proliferation of Social Networking Services (SNSs), tourism-related information has begun to be disseminated and accumulated with such services. In recent years, many tourists have begun to collect tourism information using SNSs instead of consulting guidebooks or tourism websites. The use of SNSs allows many users to share information through posts. It is not uncommon for tourists to select tourist destinations using that shared tourism information. Twitter [1], an SNS that allows the sharing of tourism information, is widely used for posting and viewing information related to events and hobbies. Posts, such as text, photographs, and images are designated as "tweets." Additionally, users can add location information to tweets at their discretion. These tweets with location information are called "geotagged tweets." Geotagged tweets are used to share what is happening where at the moment. Therefore, tweet information obtained from geotagged tweets reflects real-world circumstances. For this reason, geotagged tweets are anticipated for use as a social sensor that allows tourists to estimate and obtain tourism information for a region in real time, irrespective of location.

Sakura, or cherry blossoms, are an important Japanese tourist resource. During the cherry blossoming season, many people in Japan enjoy viewing the cherry blossoms in areas where they live. Additionally, places known as cherry blossom spots attract visitors from afar, thereby producing a bustling atmosphere. The gathering of many people engenders the dissemination of much information about various regions and spots, which can then be analyzed to estimate the best time to view the cherry blossoms. Knowing the best time to view the cherry blossoms therefore has a considerably important effect on the tourism industry and on the regional economy. In fact, people working in lodging and tourism industries, at local shops and hotels, can use information about the best time to view cherry blossoms several weeks to months in advance to forecast demand and to anticipate guests. The timing of cherry blossoms also varies because of weather conditions and climate change. The Japan Meteorological Agency [2] conducts annual observations of biological seasons, including the days of cherry blossom blooming and full bloom. However, because of difficulties in maintaining observation targets and securing personnel, the scale of biological season observations has been reduced. In 2022, the number of observation targets for plants and animals has decreased to just six, down from 34 in 2021. Although observations of the days of cherry blossom blooming and full bloom are still continuing, it is possible that they will eventually be scaled back. Therefore, using Twitter's geotagged tweets to estimate the best time to view cherry blossoms is a low-cost and sustainable method that can be done without human labor. Furthermore, estimating the best time to view cherry blossoms can provide useful information for the tourism industry and regional economy. It might also be helpful for predicting the best time to view cherry blossoms considering weather conditions and climate change.

Research on extracting tourist information from SNS can include the following. Mizutani et al. [3] developed a system that takes into account user needs and which conducts group tourism activities using SNS user information. Wang et al. [4] proposed a method of generating a tourist map using the popularity and satisfaction of tourist spots calculated from SNS-posted photos. Guanshen et al. [5] developed a tourist recommendation system that recommends tourist spots to users who want to plan a trip during specific periods, such as early autumn or Christmas holidays, using Twitter and

 ISBN: 978-1-68558-038-4

Wikipedia. As in earlier research, a method for estimating the best time to see cherry blossoms using geotagged tweets with moving averages has been proposed. Endo et al. [6] proposed a method of estimating the best time to see cherry blossoms using a simple moving average of geotagged tweets. The method is particularly useful in prefectures and municipalities where geotagged tweets are obtainable. Takahashi et al. [7] used weighted moving averages to estimate the best time to view cherry blossoms. Specifically, using a 5-day weighted moving average and a 7-day weighted moving average, they were able to improve the estimation accuracy. They performed estimation for each prefecture. However, their method showed that the prediction accuracy decreased for prefectures with fewer obtained tweets. Moreover, their method was not useful to estimate the best time to view cherry blossoms in regions where they were unable to see a trend because only a few tweets were collected per prefecture. Therefore, by increasing the number of tweets used for estimation, one can improve the prediction accuracy for prefectures that had previously been predicted to be less accurate using earlier methods. Also, one can produce estimates for prefectures that were not able to be estimated earlier because of a lack of tweets. As a method to increase the number of tweets used for estimation, the authors propose the use of words that co-occur with "cherry blossoms." Using such tweets that show the same trend as that for annual cherry blossoms, one can produce estimates even for prefectures with few tweets. As a first step, the authors report a method to increase the dataset used for cherry blossom estimation using co-occurring words.

This paper is organized as follows: In Section 2, the proposed method is explained in detail to increase the amount of data used for estimating the timing of cherry blossoms. The proposed method extracts words that co-occur with cherry blossoms, collects tweets using the extracted co-occurring words, and uses the collected tweets to estimate the timing of cherry blossom viewing. The next three sections show the results of determining the co-occurring words using the proposed method and the results of estimating the best time to see cherry blossoms. The final section, Section 4, presents the conclusion of this paper.

II. PROPOSED METHOD

This section presents a method for determining the optimal time to view certain objects of keywords, such as cherry blossoms, through the analysis of co-occurring words and collected tweets. First, we perform a co-occurrence judgment of the chosen keyword. Next, we gather tweets that include the selected co-occurring words. Finally, we estimate the optimal time for the original keyword based on the collected tweets. For this study, we particularly examine keywords that are relevant to tourism information, specifically targeting cherry blossoms. Other keywords that could be considered might be hydrangea, autumn leaves, and wildflowers. However, we have not examined those keywords. The computer platforms and environments used in making the predictions are listed in Table 1. This study specifically examines cherry blossoms. The specific steps of this method are the following.

TABLE I. USAGE ENVIRONMENT

CPU	Intel Core i5-7400 3.00GHz
Memory Capacity	24GB
Disk	TOSHIBA DT01ABA100V 1TB
GPU	NVIDIA GeForce GTX 1050
OS	Windows 10 Home
Used Language	Python 3.9.1

1. We propose a method for determining the optimal time to view a seasonal object as a keyword (in this case, "sakura" or cherry blossoms) by collecting tweets that include the keyword and using co-occurring words to infer the optimal time for viewing. The tweets were collected from February 2015 through May 2022. The keyword used for the study was "sakura," written either in kanji, katakana, or hiragana.

2. The Japanese morphological analysis tool MeCab [8] was used to conduct morphological analyses of the tweets collected in 1. MeCab is an open-source tool for performing morphological analysis of Japanese. Morphological analysis refers to the process of breaking down text into words or parts of speech. MeCab is characterized by its ability to perform high-precision analysis that incorporates evaluation of parts of speech and inflection types. It is used widely for Japanese morphological analysis. Using MeCab, the tweet text was broken down into words. The following preprocessing was performed in preparation for morphological analysis. First, cleaning was applied to the text to remove noise, such as URLs and symbols. Next, normalization was done to facilitate analysis of the words. Because Japanese tweets include a mix of full-width and half-width characters, full-width numbers were converted to half-width numbers, half-width katakana was converted to full-width katakana, alphabetical characters were converted entirely to lowercase, and numbers were all replaced with "0". After this preprocessing, MeCab was used to break down the tweet text into words. Additionally, MeCab and IPAdic, a dictionary for morphological analysis, were used to extract words corresponding to the three parts of speech that are likely to be used as collocates: "noun," "adjectival noun," and "verb." Then, the frequency of each was counted. Here, "adjectival nouns," which represent the stem of an adjective, are a part of speech introduced as a major classification by UniDic [9].

3. We conducted co-occurrence judgment of the top 1% of the words collected in "2". Words found to have frequency of less than 1% were determined to have low relevance to the keyword "sakura" and were not explored as co-occurring words because their occurrence in tweets including "sakura" from February 2015 through May 2022 was in the single digits. For the method of co-occurrence judgment, we calculated the skewness and kurtosis of the frequency distribution of each word and determined the words which met the criteria as co-occurring words. Skewness and kurtosis are statistics that are used to indicate

how much a distribution deviates from a normal distribution and how sharp it is. They are regarded as appropriate indicators for finding co-occurring words with the same peak as the keyword during the annual tweet trend. The criteria for determination are given as (1) and (2) below. The formulas respectively represent skewness and kurtosis of the frequency distribution of each word as *S* and *K* , and the skewness and kurtosis of the keyword "sakura" as *Ss* and *Ks*, respectively.

$$S > Ss \quad (1)$$

$$(Ks - 1) \leq K \leq (Ks + 1) \quad (2)$$

In this section, the method used to calculate skewness and kurtosis is described. The skewness and kurtosis were calculated for tweets that include the word in question in the text, extracted between January 1, 2018 and December 31, 2018. The skewness (*Ske*) and kurtosis (*Kur*) are calculated using the formulas shown respectively in (3) and (4) below, where *n* represents the number of days in the specified period, x_i denotes the value for the *i*th day (i = 1, 2, ..., n), x-bar stands for the mean value for the specified period, and s represents the sample standard deviation. In this case, n is 365 days because the period used for this study is January 1, 2018 through December 31, 2018.

$$Ske = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3 \quad (3)$$

$$Kur = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^4}{s^4} - \frac{3(n-1)^2}{(n-2)(n-3)} \quad (4)$$

To confirm the trend of tweets showing the same tendency as that of cherry blossoms throughout the year, it was necessary to set the period to include all months from January through December. Additionally, while the tweets were collected by our research team on a server that we manage, there were periods during which the collection was stopped because of power outages. To avoid any skewing of the calculated skewness and kurtosis values, it was necessary to minimize the inclusion of periods when collection was not possible. Therefore, we adopted the period from January 1, 2018 through December 31, 2018, which satisfies these two conditions, for the calculation of skewness and kurtosis.

4. To ascertain the peak period of cherry blossoms, tweets including the words identified as co-occurring words in "2" were collected. The peak period was estimated from the collected tweets. Estimation of the peak period was conducted using the same moving average method as the conventional method. The moving average used was a simple moving average. The *n*-day simple moving average is obtainable from (5) as shown below.

$$\text{Avg } n = \left(x_i + x_{i-1} + \cdots + x_{i-(n-2)} + x_{i-(n-1)}\right)/n \quad (5)$$

The next two (6) and (7) were used to find the peak period by judging the period that simultaneously satisfies both equations for more than three consecutive days. Equation (6) identifies the period during which the number of tweets is higher than the average for a year. Equation (7) shows the period during which the number of tweets increases suddenly. The peak period is therefore identified by combining these two equations. In (6) and (7), x_i represents the number of tweets on day *i*. Also, Avg *n* represents the moving average over *n* days.

$$x_i > \text{Avg } 365 \quad (6)$$

$$\text{Avg } 10 < \text{Avg } 20 \quad (7)$$

The traditional method only collected tweets that included the word "sakura", resulting in an approximate daily tweet count of 30. However, the use of co-occurring words in the text caused an average of 400 tweets, which rendered the slight changes identified using the traditional moving-average method unreliable and which markedly reduced the estimation accuracy. Therefore, a moving average that was adjusted to the number of tweets including co-occurring words was used.

For this study, we specifically examined the increase of the data volume, which has persisted as an important difficulty associated with the conventional method. Moreover, we sought to solve it merely by changing the data collection method. However, because of differences in the data volume, it became necessary to change the conditions used for estimation. Finally, we adjusted them to match the results we sought.

5. Finally, we compared the estimated periods obtained using the conventional method and the estimated the period obtained only from co-occurring words. We found that both results had similar accuracy. Therefore, we thought that an increase in accuracy could be expected by combining these two estimation results.

III. RESULTS AND DISCUSSION

This paper presents results of our co-occurrence word determination and visual appeal estimation. The results obtained for the co-occurrence words collected for each of the three prefectures of Tokyo, Kyoto, and Shizuoka are presented respectively in Tables 2, 3, and 4. Table 2 for Tokyo shows the presence of the name of the famous Ueno Onshi Park, known for its cherry blossoms, as well as words related to cherry blossoms. Similarly, Table 3 for Kyoto shows the presence of the names Gion and Uji, which are also famous cherry blossom spots. Table 4 for Shizuoka shows the famous cherry blossom spots of Kawazu and Suruga. From these data, it can be inferred that the proposed method of co-occurrence word determination can be used to find relevant words. Additionally, it is noteworthy that Tokyo was found to have the most co-occurrence words among the three prefectures. That finding can be attributed to the fact that the total number of tweets collected for Tokyo was 179,332, whereas the numbers of tweets collected for Shizuoka and Kyoto were 15,596 and 16,790: the tweets collected for Tokyo were about 10 times more numerous than for either of the other two prefectures.

Figures 1–6 present results obtained when estimating the optimal viewing period for cherry blossoms from March 1 through April 30 of 2022. The correct period shown in the figures refers to the period from the date of flowering announced by the Meteorological Agency through the date of full blooming. That period is used for evaluation of the estimated results. In each figure, the number of tweets including the word "cherry blossoms" is shown in a stacked bar graph. The estimated optimal viewing period is shown in a shaded area. Figure 1 shows the period estimated as the optimal viewing period for Tokyo, either by the conventional method or by the proposed method. Figure 2 presents the

period estimated as the optimal viewing period for Tokyo by both the conventional method and the proposed method. Figure 3 shows the period estimated as the optimal viewing period for Kyoto, either by the conventional method or by the proposed method. Figure 4 shows the period during which both the conventional method and the proposed method estimated the best viewing time for Kyoto prefecture. Figure 5 shows the period during which either the conventional method or the proposed method estimated the best viewing time for Shizuoka prefecture. Figure 6 shows the period during which both the conventional method and the proposed method estimated the best viewing time for Shizuoka prefecture. As Figures 1, 3, and 5 show, the estimated period for all three prefectures is longer than the actual period, which is likely attributable to the fact that the actual period is from the blooming day through the full blooming day. Actually, it takes about 10 days to 2 weeks from the full blooming day until cherry blossoms start to scatter [10]. Early blooming cherry blossoms known as Kawazuzakura, which are famous for blooming in early January and reaching their best viewing time by early March, are thought to be related to estimation of the best viewing time from March 1 [11], as shown in Figure 5. This finding is consistent with the period stated as tourist information. Moreover, this finding is thought to have led to estimation of the best viewing time as that from early March.

We evaluated results of estimating the blooming period of cherry blossoms using both traditional methods and methods that use co-occurring words. Table 5 presents results obtained from the evaluation using the metrics of

TABLE II. LIST OF CO-OCCURRING WORDS IN TOKYO PREFECTURE (IN JAPANESE)

cherry blossom	sakura	Ueno park	さく (bloom)	ソメイヨシ (Yoshino cherry)
ヒルズ (hills)	ミッドタウン (midtown)	七 (seven)	五反田 (Gotanda)	今年 (this year)
付着 (adhesion)	六義 [garden name]	初 (first time)	咲い (bloom[missing words])	咲き (bloom[missing words])
国際 (international)	大学 (university)	大崎 [place name]	始め (start)	川沿い (river side)
恩賜 [park name]	日和 (weather)	日野 [place name]	最後 (last)	本髪 (hair)
来年 (next year)	毎年 (every year)	江東 [place name]	皆 (everyone)	花 (flower)
花びら (petal)	花園 [place name]	見物 (sightseeing)	見頃 (best time to see)	調布 [place name]
開花 (blooming)	靖国 [name of shrine]			

TABLE III. LIST OF COLLOCATIONS IN KYOTO PREFECTURE (IN JAPANESE)

cherry blossom	割 (divide[missing words])	咲き (bloom[missing words])	宇治 [place name]	平野 [place name]
年 (year)	散っ (scatter[missing words])	淀川 (Yodo rever)	疏水 [place name]	祇園 [place name]

TABLE IV. LIST OF CO-OCCURRING WORDS IN SHIZUOKA PREFECTURE (IN JAPANESE)

みなみ (south)	並木 (row of trees)	分 (divide[missing words])	咲い (bloom)	咲き (bloom)
始め (start)	平 [place name]	新聞 (newspaper)	早咲き (early bloom)	梨子 [character name]
河津 [place name]	清水 [place name]	菜の花 (rape blossoms)	駿府 [place name]	
神社 (shrine)	蹴上 [name of shrine]	開花 (blooming)		

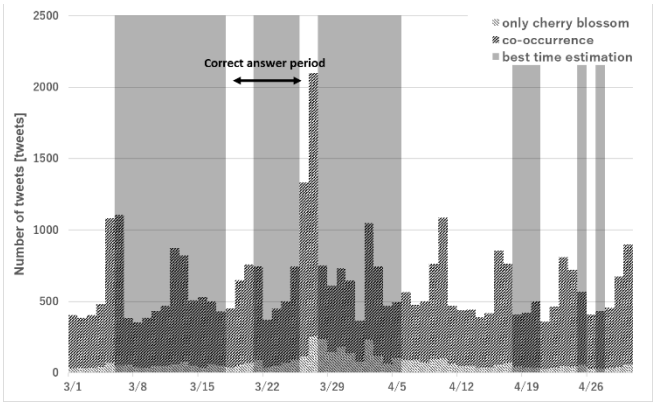

Figure 1. Estimated results for best viewing time in Tokyo prefecture (or).

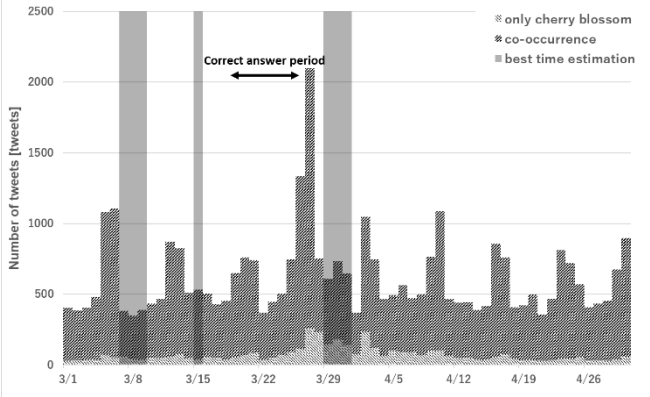

Figure 2. Estimated results for best viewing time in Tokyo prefecture (and).

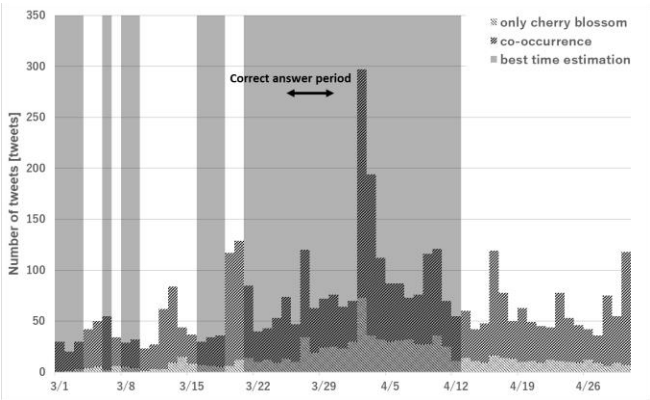

Figure 3. Estimated results for best viewing time in Kyoto prefecture (or).

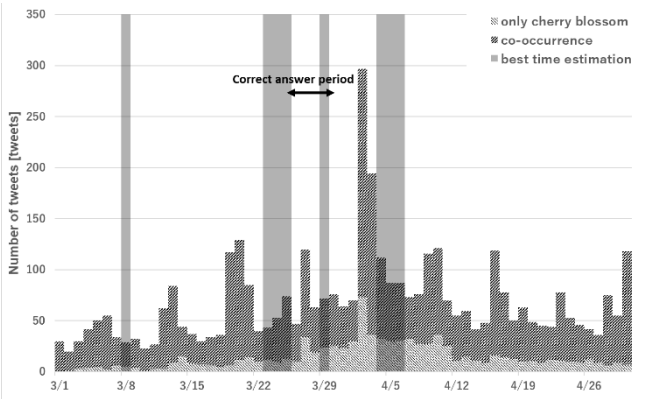

Figure 4. Estimated results of the best viewing time in Kyoto prefecture (and).

precision and recall, which are indicators of the accuracy of estimation. The Japan Meteorological Agency defines the blooming period as the period from the start of blooming to full bloom. The recall metric is used to evaluate how well the estimated period falls within this period. The precision metric is calculated as the proportion of the estimated period which falls within the blooming period. In Table 5, "sakura" refers to results of estimation conducted using traditional methods, whereas "co-occurring words" refers to results of

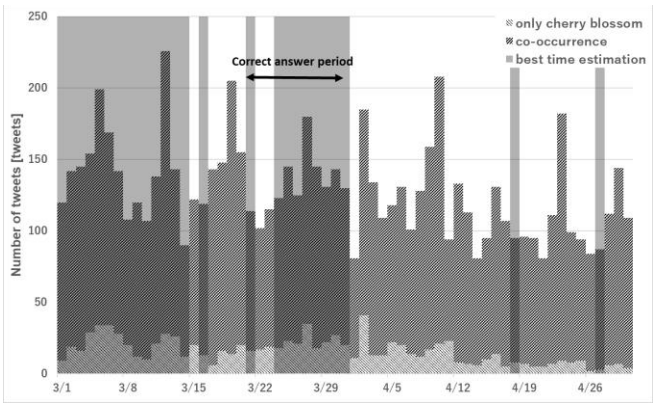

Figure 5. Estimated results for best viewing time in Shizuoka prefecture (or).

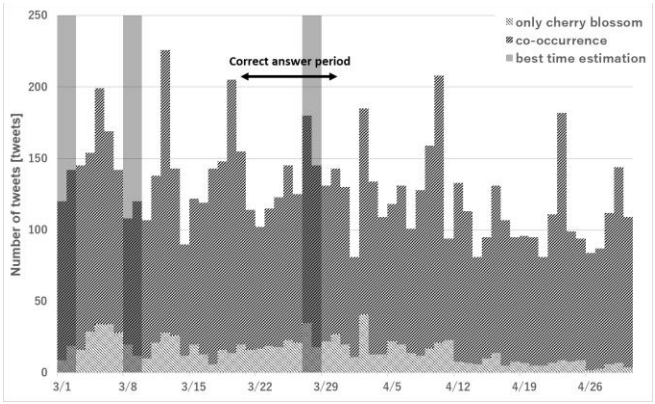

Figure 6. Estimated results for best viewing time in Shizuoka prefecture (and).

estimation using the proposed method. The "or" and "and" columns respectively refer to the evaluation of periods estimated as blooming by either the traditional method or the proposed method, and by both methods.

As Table 5 shows, when doing estimation with co-occurring words in Tokyo and Shizuoka, the prediction accuracy is lower than when estimation is done using traditional methods. In contrast, when using co-occurring words to conduct an estimation for Kyoto, the prediction accuracy is higher than when doing an estimation using traditional methods. Therefore, the results of traditional methods and the proposed methods vary depending on the prefecture. However, when evaluating the results of estimating the period during which either traditional methods or proposed methods are regarded as the best time, the reproducibility of the three prefectures can be improved. Therefore, the accuracy can probably be improved by combining traditional methods and the proposed methods. In addition, since the proposed method increases the amount of data by using co-occurrence word judgments, the amount of computation required to estimate the best time to visit is larger than that of the prior method. Since the objective of the present study is to suppress the decline in prediction accuracy when the amount of data used for estimating the best time to visit is reduced, which was an issue with the

TABLE V. RESULTS OF BEST VIEWING EVALUATION

Method	Prefecture	Recall (%)	Precision (%)
sakura	Tokyo	50.0	26.7
co-occurring words		12.5	4.3
or		52.5	16.1
and		0.0	0.0
sakura	Kyoto	57.1	23.5
co-occurring words		85.7	26.1
or		100.0	21.9
and		42.9	37.5
sakura	Shizuoka	70.0	50.0
co-occurring words		30.0	16.7
or		80.0	30.8
and		20.0	33.3

prior method, the large amount of computation is not taken into account. It should be noted that although the results do not contradict the objective, the increase in accuracy was obtained in exchange for the increase in computational load.

IV. CONCLUSION AND FUTURE WORK

This report described our study of a method for estimating the cherry blossom blooming period using geo-tagged tweets and co-occurring words. To achieve higher accuracy of estimation, we proposed a method for increasing the data amount using co-occurring words related to keywords. When estimating the cherry blossom blooming period using co-occurring words, we demonstrated a prediction accuracy equivalent to that achieved using the conventional method, which used only cherry blossoms. However, the estimated period differed from that of the conventional method. Using this difference, we were able to improve the accuracy by combining results of the conventional method and the proposed method. No improvement in accuracy was found when using only co-occurring words for prediction, perhaps because tweets including words that co-occur with "cherry blossoms" also included many tweets that were unrelated to the plant. Moreover, these noise tweets might have impaired the method, making it less accurate. Removal of these noisy tweets is left as a task for future work. In future studies, we would like to perform estimation for seasonal words that are unrelated to cherry blossoms. Thereby, we can ascertain whether the proposed method is effective for other words.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number JP22K13776, JP20K12081, and the Okawa Foundation Research Grant.

REFERENCES

[1] Twitter. [Online]. Available from : https://twitter.com/ [retrieved: 1, 2023]

[2] Japan Meteorological Agency. [Online]. Available from: https://www.jma.go.jp/jma/indexe.html [retrieved: 1, 2023].

[3] Y. Mizutani and K. Yamamoto, “A sightseeing spot recommendation system that takes into account the change in circumstances of users,” ISPRS International Journal of Geo-Information, vol. 6, no. 10, pp. 303, 2017.

[4] Y. Wang and W. Yue, “Proposal and Evaluation of a Pictorial Map Generation Method Based on the Familiarity and Satisfaction of Sightseeing Spots Calculated from Photos Posted on SNS (in Japanese),” Journal of the Japan Personal Computer Application Technology Society, vol. 16, no. 1, pp. 1–10, 2021.

[5] G. Fang, S. Kamei, and S. Fujita, “How to extract seasonal features of sightseeing spots from Twitter and Wikipedia (preliminary version),” Bulletin of Networking, Computing, Systems, and Software, vol. 4, no. 1, pp. 21–26, 2015.

[6] M. Endo, M. Takahashi, M. Hirota, M. Imamura, and H. Ishikawa, “Analytical Method using Geotagged Tweets Developed for Tourist Spot Extraction and Real-time Analysis,” International Journal of Informatics Society (IJIS), vol. 12, no. 3, pp. 157–165, 2021.

[7] M. Takahashi, M. Endo, S. Ohno, M. Hirota, and H. Ishikawa, “Automatic detection of tourist spots and best-time estimation using social network services,” International Workshop on Informatics 2020, pp. 65–72, 2020.

[8] T. Kudo, Mecab: Yet another part-of-speech and morphological analyzer. [Online]. Available from :http://mecab.sourceforge.net/ , [retrieved: 7, 2022].

[9] National Institute for Japanese Language and Linguistics. Electronic Dictionary with Uniformity and Identity . [Online]. Available from: https://clrd.ninjal.ac.jp/unidic/ [retrieved: 1, 2023].

[10] Spring Season, What Month do Cherry Blossoms Bloom and When Do They Fall? Survey of cherry blossom season by prefecture . [Online]. Available from: https://nihonail.com/season/2195.html [retrieved: 1, 2023].

[11] Kawazuzakura. [Online]. Available from: https://hanami.walk-erplus.com/special/kawazu/ [retrieved: 1, 2023].

 ISBN: 978-1-68558-038-4