Utilization of Information Interpolation using Geotagged Tweets

Masaki Endo
Polytechnic University
Tokyo, Japan
e-mail endou@uitec.ac.jp

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

Hiroshi Ishikawa
Tokyo Metropolitan University
Tokyo, Japan
e-mail ishikawa-hiroshi@tmu.ac.jp

ABSTRACT

Along with the spread of social media, it has become possible to extract events occurring in the real world in real time. A benefit of analysis using data with position information is that it can accurately extract an event from a target area to be analyzed. However, because social media data include few data with location information, the amount to analyze is insufficient for almost all areas: we cannot fully extract most events. Therefore, efficient analytical methods must be devised for the accurate extraction of events with position information, even in areas with few data. For this study, we use geotagged tweets along with interpolation to estimate the best time to observe biological seasons when doing sightseeing such as cherry-blossom and autumn leaf viewing in areas and sightseeing spots. Herein, we explain the analysis results obtained using information interpolation and analysis of cherry blossoms in Japan during 2017.

Author Keywords

trend estimation; phenological observation; Twitter

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

INTRODUCTION

Because of the wide dissemination and rapid performance improvement of various devices such as smart phones and tablets, diverse and vast data are generated on the web. Particularly, social networking services (SNSs) have become popular because users can post data and various messages easily. Twitter [1], 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.

We specifically consider tourist information provision using real-time information from Twitter. According to a survey reported in the Inbound Landing-type Tourism Guide [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, the information update frequency of that medium is low. Because each local government, tourism association, and travel company independently provides information about travel destination locales, it is difficult for tourists to collect information for "now" tourist spots. Therefore, providing current, useful, real-world information for travelers by capturing changes of information in accordance with the season and relevant time period of the tourism region is important for the travel industry.

Tourist information for best times requires a peak period, which means that the best time is neither a period after or before falling flowers, but a precisely defined period to view blooming flowers. Furthermore, the best times differ among regions and locations. Therefore, for each region and location, it is necessary to estimate the best time for phenological observations. Estimating best-time viewing periods requires the collection of large amounts of information having real-time properties. For this study, we use Twitter data obtained from many users throughout Japan. We use Twitter, a typical microblogging service, and also use geotagged tweets that include position information sent in Japan to ascertain the best time (peak period) for biological season observation by region. The geotagged tweets are useful as social indicators reflecting real-world circumstances. They are a useful resource supporting a real-time regional tourist information system in the tourism field. Therefore, our proposed method might be an effective means of estimating the best time to view events other than biological seasonal observations.

To analyze information of each region from Twitter data, it is necessary to specify a location from tweet information. Geotagged tweets can identify places. Therefore, they are effective for analysis. However, because geotagged tweets account for only a very small proportion of the total information content of tweets, it is not possible to analyze all regions. We propose a method to provide tourists with information about sightseeing spots after processing a small amount of data in real time by interpolation using geotagged tweets.

RELATED WORKS

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

UISTDA '18, March 11, Tokyo, Japan.

analyzing large amounts of digital data. Some studies have been conducted to predict real world phenomena using large amounts of social big data. Phithakkitnukoon et al. [3] analyze details of traveler behavior using data from mobile phone GPS location records such as embarkation places, destinations, and traveling mode on a personal level. Mislove et al. [4] develop a system that infers a Twitter user's feelings from tweet text and which visualizes changes of emotion in space–time. Based on research to detect events such as earthquakes and typhoons, Sakaki et al. [5] propose a method to estimate real-time events from Twitter tweets. Cheng et al. [6] estimate Twitter users' geographical positions at the time of their contributions, without the use of geotags, by devoting attention to the geographical locality of words from text information in articles posted on Twitter. Yamada [7] analyzes Japanese blog data and proposes a method to identify seasonal words using simple autocorrelation analysis. Krumm et al. [8] propose a method to detect events using time series analysis of geotagged tweet volumes from localized areas. Various studies have analyzed spatiotemporal data, but research to estimate viewing periods using interlinkage is a new field.

OUR PROPOSED METHOD

We describe the best-time estimation method of organisms by analysis using geotagged tweets that include organism names. Best-time estimation, as defined for this paper, is estimation of the period during which creatures at tourist spots are useful for sightseeing. Such information can be useful reference information when visiting tourist spots. It supports estimation of the period during which a tourist can enjoy the four seasons by viewing cherry blossoms and autumn leaves. However, geotagged tweets are far fewer than tweets without geotags. For that reason, although it is possible to estimate the best time in a prefecture unit or municipality, finely honed analyses have been impossible. Nevertheless, the best time to visit sightseeing spots can be estimated with finer granularity using the method with interpolation proposed in this paper.

In the following subsections, we describe the collection of geotagged tweets to be analyzed, character preprocessing for conducting analysis, and interpolation using Kriging.

Data collection

This section presents data collection. Geotagged tweets sent from Twitter are a collection target. The range of geotagged tweets includes the Japanese archipelago (120.0°E – 154.0°E, and 20.0°N – 47.0°N) as the collection target. The collection of these data was done using a streaming API [9] provided by Twitter Inc.

Next, we describe 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: this is a rare characteristic for text data. However, the geotagged tweets we collected are an average of 500 thousand tweets per day. We use about 250 million geotagged tweets from 2015/2/17 through 2017/5/13. In addition, although the author has a Twitter account, which is necessary to use the API, the author never tweets. Moreover, even for personal account holders, tweets do not include location information in some areas. Therefore we do not consider excluding the tweets of the author's own account. Using these data, we calculated the best time for flower viewing, as estimated using the processing described in the following sections.

Preprocessing

This section presents preprocessing. Preprocessing includes reverse geocoding and morphological analysis, as well as database storage for data collected through the processing described in the previous subsection.

From latitude and longitude information in the individually collected tweets, reverse geocoding is useful to identify prefectures and municipalities by town name. We use a simple reverse geocoding service [11] available from the National Agriculture and Food Research Organization in this process.

Morphological analysis divides the collected geotagged tweet morphemes. We use the "Mecab" morphological analyzer [12].

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

Interpolation using Kriging

This section presents the method of interpolation, for which we used Kriging [13], an estimation method used for estimating values for points where information was not acquired. It is impossible to estimate from the number of geotagged tweets of each sightseeing spot when conducting detailed analysis at each sightseeing spot. Therefore, geotagged tweets that have seven significant digits and the same latitude and longitude information are judged to have originated from the same spot. As an example, the tweet's position information of (latitude, longitude) = (34.93162536621094, 135.72979736328125) is truncated to (latitude, longitude) = (34.93162, 135.72979). Then, the tweets from the same point were counted for each date. Furthermore, by dividing the total for each point by the total number of tweets on each day, we calculated the weight of each spot.

We attempted estimation by interpolation using data aggregated for each spot. The estimated value of the target data at a point S_0 is shown in formula (1) as a weighted average of the measured values $Z(S_i)$ (i = 1, 2..., N) at N points S_i around point S_0. Then we assigned value Z to tweets including the target word and Z. Here, N represents the 30 nearby targeted tweets. λ denotes a spherical model with decreased influence as distance increases. As described in this paper, weighting is done only by the number of tweets existing in the same spot. However, consideration of the weight of the tweet itself, such as using the number of retweets to tweets, is also necessary for future studies.

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

$Z(S_i)$:Measurement value at i-th position

λ_i :Unknown weighting of measured value at i-th position

S_0 :Predicted position

N :Number of measurements

EXPERIMENTS

In this section, we explain the experiment for information interpolation for cherry blossoms in 2017, using the method described in the previous section. We are conducting estimation experiments for cherry blossoms, autumn leaves, and other phenomena from 2015. As described herein, we used the period of cherry blossoms in 2017 while studying the interpolation method to improve estimation accuracy. The following subsections describe experimental datasets.

Datasets

Datasets used for this experiment were collected using streaming API, as described for data collection. The data, which include about 250 million items, are geotagged tweets from Japan during 2015/2/17 – 2017/5/13. The estimation experiment conducted to ascertain the best-time viewing of cherry blossoms uses the target word "cherry blossom," which is "桜", "さくら", and "サクラ" in Japanese. We analyzed tweet texts that include the target word. About 100,000 tweets during the experiment period included the subject word.

The subject of the experiment was set as tourist spots in Tokyo. In this report, we describe "Takao Mountain," "Showa Memorial Park," "Shinjuku Gyoen," and "Rikugien." Figure 1 presents the target area locations. A, B, C, and D in the figure respectively denote "Takao Mountain," "Showa Memorial Park," "Rikugien," and "Shinjuku Gyoen." In this experiment, about 30,000 tweets including the target word in Tokyo were found. In this experiment, all tweets made by the same user are also used as targets for analysis if they are tweets including the target word.

We conducted experiments of the following two kinds using these datasets. The first is an experiment using the number of tweets including the target word and the sightseeing spot name without interpolation. This experiment was compared as Baseline to confirm the usefulness of interpolation proposed in this paper. The second is an experiment using interpolation. In this experiment, we used Kriging in the earlier section.

We present the example of Mt. Takao in Hachioji city. Figure 2 shows cherry-blossom-related tweets in Hachioji city from 2017/1/1 to 2017/5/13. The point denoted by A is Mt. Takao. Few geotagged tweets are related to cherry blossoms near Mt. Takao. For that reason, one cannot estimate the best time merely using tweets from A.

Therefore, for this study, we interpolated the amount of information by Kriging using information of Hachioji city with Mt. Takao. Interpolation is done on a daily basis. The experiment results are presented in the next subsection.

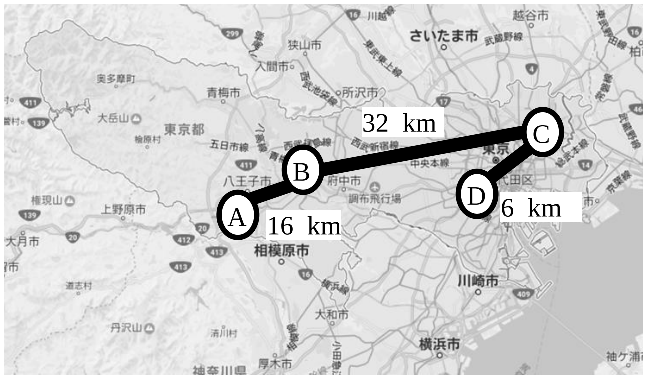

Figure 1. Location of the target area.

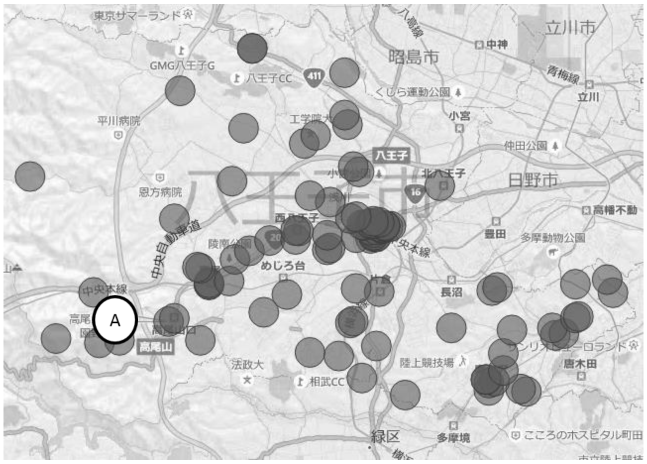

Figure 2. Positions of targets.

Experimental results

This section presents experimentally obtained results for estimating the best time. Figure 3 presents results for the estimated best-time viewing in 2017 using the target word 'cherry blossoms' in the target tourist spots. The black part of the figure represents the number of tweets containing a target word and sightseeing spot name. The light gray part represents best-time viewing as determined using the proposed method of interpolation.

At tourist spots targeted for the experiment in 2017, as portrayed in the black part of Figure 3, many data were obtained for B, C, and D. The maximum number of tweets per day was about 20. These results confirmed that some estimation can be accomplished without interpolation.

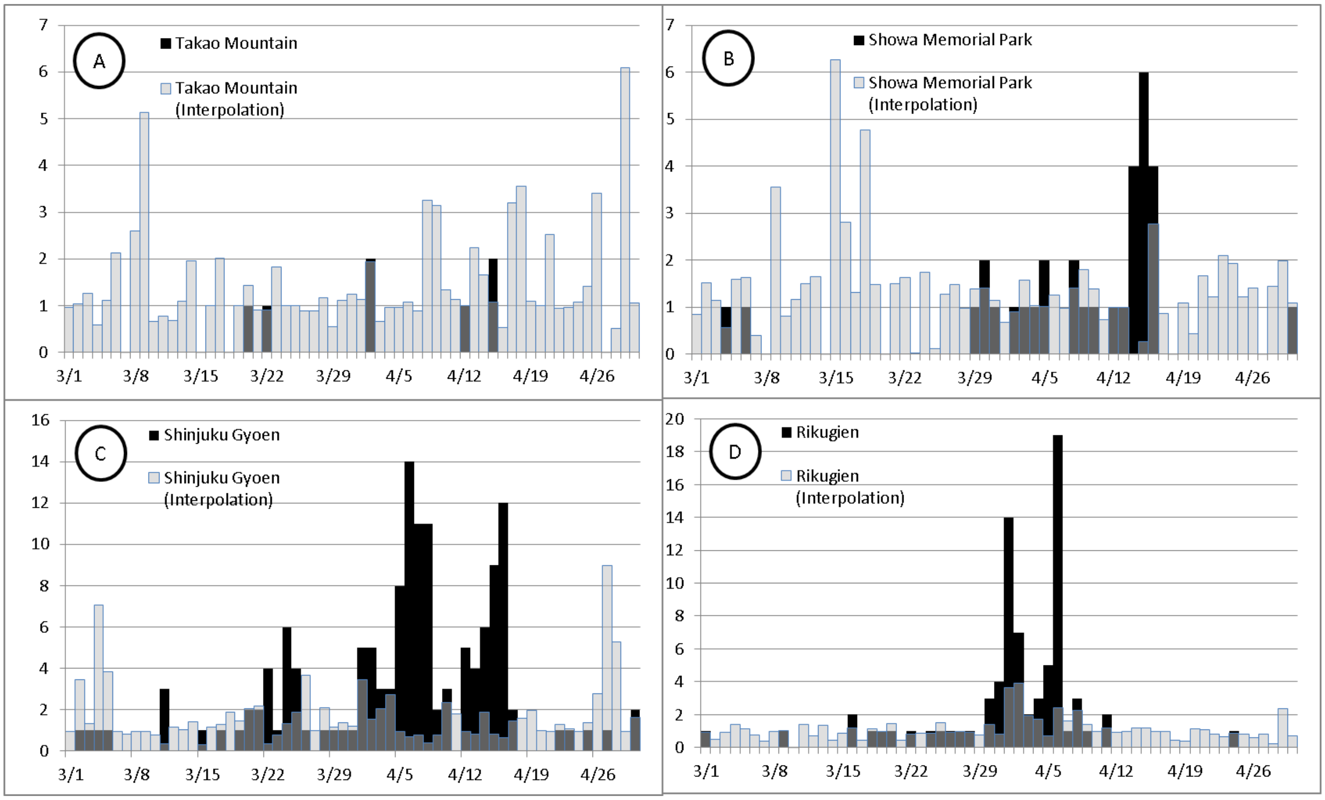

Figure 3. Experimental results.

The light gray part of Figure 3 portrays an experimentally obtained result from interpolation results including the tourist spots. Apparently, A was able to produce an estimate using the proposed method by increasing the number of tweets using interpolation with surrounding tweets. For B and C, the useful information was included in the tweet not co-occurring with the tourist spot name. Therefore, we confirmed interpolation related to other kinds of cherry blossoms in early March of B and C and late April of C. In addition, for D, there are days when it can be determined more accurately by interpolating the number of tweets.

Therefore, these results confirmed the possibility of estimating the peak period, even for an area without tweets, using data interpolation and overall tweet number interpolation.

CONCLUSION

As described in this paper, we proposed an interpolation method to improve the accuracy of tourism information related to phenological observations. The results of the cherry blossom experiment conducted at the tourist spots in Tokyo in 2017 confirmed the trend of improved estimation accuracy using information interpolation. We confirmed the possibility of applying this proposed method to the estimation of viewpoints and sightseeing spots with few tweets. However, in regions with no geo-tagged tweets, another method must be considered. Research can be conducted in the future to verify whether similar results are obtained for other biological seasonal observations.

ACKNOWLEDGMENTS

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

REFERENCES

1. Twitter. It's what's happening. 2017. Retrieved September 2, 2017 from https://Twitter.com/

2. Ministry of Economy. Trade and Industry. Inbound Landing-Type Tourism Guide. 2017. Retrieved November 10, 2017 from http://www.mlit.go.jp/common/001091713.pdf (in Japanese).

3. S. Phithakkitnukoon, T. Teerayut Horanont, A. Witayangkurn, R. Siri, Y. Sekimoto, and R. Shibasaki. 2015. Understanding tourist behavior using large-scale mobile sensing approach: A case study of mobile phone users in Japan. *Pervasive and Mobile Computing Volume* 18: 18-39.

4. A. Mislove, S. Lehmann, Y.Y. Ahn, J-P. Onnela, and J. Niels Rosenquist. Understanding the Demographics of Twitter Users. 2011. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (icwsm 2011), 554–557.

5. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. 2010. *WWW 2010*, 851–860.

6. T. Kaneko and K. Yanai. Visual Event Mining from the Twitter Stream. 2010. WWW '16 Companion, In *Proceedings of the 25th International Conference Companion on World Wide Web*, 51–52.

7. K. Yamada. Detecting two types of seasonal words using simple autocorrelation analysis. 2017. IEEE Big Data 2017 Workshops. In *Proceedings of the Second International Workshop on Application of Big Data for Computational Social Science.*

8. J. Krumm and E. Horvitz. Eyewitness: identifying local events via space-time signals in Twitter feeds. 2015. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '15). ACM*, New York, NY, USA,, Article 20, 10 pages. DOI: https://doi.org/10.1145/2820783.2820801

9. Twitter Developers. Twitter Developer official site. 2017. Retrieved April 2, 2017 from https://dev.twitter.com/

10. Y. Hashimoto and M. Oka. Statistics of Geo-Tagged Tweets in Urban Areas (<Special Issue>Synthesis and Analysis of Massive Data Flow). 2012. *JSAI* vol. 27, 4: 424–431 (in Japanese).

11. National Agriculture and Food Research Organization. Simple reverse geocoding service. 2017. Retrieved November 18, 2017 from https://www.finds.jp/rgeocode/index.html.ja

12. MeCab. Yet Another Part-of-Speech and Morphological Analyzer. 2017. Retrieved November 10, 2017 from http://taku910.github.io/mecab/

13. M. A. Oliver. Kriging: A Method of Interpolation for Geographical Information Systems. 1990. *International Journal of Geographic Information Systems* vol. 4: 313–332.