Regular Paper

Estimating a Specific Position Related to an Event for Deflation Detection

Takuma Toyoshima[*], Takuo Kikuchi[**], Masaki Endo[**], Shigeyoshi Ohno[**] and Hiroshi Ishikawa[***]

[*]ICT Production Support, Polytechnic Center Shiga, Japan
[**]Division of Core Manufacturing, Polytechnic University, Japan
{kikuchi, endou, ohno}@uitec.ac.jp
[***]Graduate School of System Design, Tokyo Metropolitan University, Japan
ishikawa-hiroshi@tmu.ac.jp

Abstract - Sensors are useful to track a person's movement at an event. Physical sensors are expensive to install and maintain. Therefore, instead of physical sensors, social network service (SNS) users can be treated as sensors to observe real-world events. To do so, many data are required. However, for information protection, few SNSs retain accurate location information. To ascertain the movement of a person at an event using a social sensor, which costs less than a physical sensor, one must infer the location information of the social sensor. Therefore, we assess a method of estimating position information related to a specific event. Estimation accuracy was evaluated using actual data from tweet messages recorded in Chiyoda ward, Tokyo. For SNSs with accurate location information, cross-validation is used to evaluate the location estimation method. However, because few SNSs have accurate location information, we also use SNSs with location information for each city, ward, town, and village. In this case, the position estimation result is considered by evaluating the accuracy of the SNS as a sensor. By estimating the position, one can increase the SNS data posted near a specific event. Furthermore, using this SNS data, one can track the movement of people at an event more accurately.

Keywords: Deflation structure, Location estimation, Real time analysis, Social sensor, Twitter

1 INTRODUCTION

Computerized devices and systems used in modern society provide many means for real-time acquisition of diverse information. One means is Twitter, a microblogging service that shares short sentences called "Tweets" of 140 or fewer characters. The service is widely used throughout the world, as it is in Japan. Many users regard it as a medium by which they can post information continually and casually. Posting of location information can be done easily via a smartphone with so-called geotags. This social medium can therefore immediately notify many people of what is happening and where. Based on these characteristics, such a social medium is anticipated for use as a social sensor for observing the real world without expensive physical sensors.

Social sensors can elucidate a situation in real time even if one is not present on the scene. For example, if one can estimate the best time to view cherry blossoms and autumn leaves before visiting a place, one could actually go to a place with no concern that cherry blossoms have not yet bloomed or that they have already fallen. If public transportation is halted and a person knows that people are crowding into stations, then that person can avoid those crowds. If a person knows that congestion in an area has eased, then the person could plan a route through the area. If a reveler wants to attend a Halloween party in Shibuya, then that person would want to hurry while the party is still exciting. After the Halloween party settles down, it might not be so interesting. Alternatively, to avoid a raucous party, one might wait until after it has settled down. Social sensors would be useful to inform people who make such choices.

Analyzing today's events using yesterday's data is not always helpful, but predicting the movements of other people in a specific place in real time can help a person decide whether to visit a certain place or not. Assuming that one is not actually present in a certain place, Twitter data can be useful to evaluate concentrations of people in real time while avoiding deployment of expensive sensors. This study was conducted to produce a means of real-time estimation of human motion by analyzing geotagged tweets.

To infer the exact movement of a person using Twitter, each tweet must have accurate time and location information. Without accurate information, any inference of a person's movement would be unreliable. Regarding time information, the tweet posting time is recorded accurately in seconds. Regarding location information, few tweets can be recorded accurately from the viewpoint of confidentiality of personal information. Therefore, position estimation is necessary. A tweet includes 140 characters of textual information. Position estimation can be achieved by analyzing the contents. Thereby, one can estimate whether a tweet was posted near an event for which one wants to know the movement of people. By this simple process, one can ascertain the movement of people at the event accurately if one can accurately assess tweets posted near the target event. The position estimation here is not intended to estimate the position more accurately. It is only necessary to be able to estimate whether or not the event for which we want to know the movement of people is near this event. The ultimate goal is to determine the movement of people, especially deflation, which indicates the spreading out of people, with a social sensor. Also, it is necessary to estimate the tweet position to improve the accuracy of the deflation detection.

2 RELATED WORK

Kleinberg [2] proposes a method for modeling text stream bursts and for extracting structures. This method is based on modeling a stream using an infinite state automaton. A salient benefit of Kleinberg's approach is that it can represent the burst duration, degree, and weight for each topic. Therefore, it is used widely for various applications. Nevertheless, it is unsuitable for real-time burst detection because analyses cannot be done immediately for occurrence of a certain event.

Studies conducted by Zhu and Shasha [3] [4] and by Zhang and Shasha [5] examine bursting algorithms that monitor bursts efficiently over multiple window sizes. These techniques enable nearly real-time burst detection by shortening of the monitoring interval. However, they require monitoring of the number of occurrences of events at regular intervals. Data must be stored even if no event has occurred.

Ohara et al. [6] propose a method for detecting the period during which burst diffusion occurs from the information diffusion series observed on social networks. They detected the burst period on Twitter, but they must build a social network to observe the spread of information related to the network.

Ebina et al. propose a method for real-time burst detection [7] [8]. The method achieves detection by inference of whether each event (each tweet posting) is a burst, or not. The number of calculations is reduced by compressing data held at the time of occurrence of concentrated events. Burst detection with high real-time capability is achieved, but it remains unclear whether the burst state continues or immediately ends solely by the burst occurrence.

Endo et al. use a moving average to make full-fledged decisions [9] [10]. This method detects burst occurrence and burst state continuation and convergence. However, because the tweet occurrence frequency is used with a fixed window size, real-time properties are quantized by the window size. Using the Tweet Posting frequency requires setting of a certain posted interval for frequency calculation. This fixed interval impairs real-time performance.

Large amounts of tweet data are necessary to estimate people's movements in real time. However, few tweets include any location information: tweets with accurate location information are even fewer. Furthermore, much location information is ambiguous. Therefore, research is underway to obtain location information from tweet contents [11] [12]. Methods have been designed to obtain location information by analyzing the vocabulary included in the tweet text. Such methods identify a target, such as an event or building, in a tweet that has no location information in the first place. Therefore, it is inferred that the tweet was tweeted from the event venue or from the position of the object. However, for the present study, it is assumed that a person has location information for each city, ward, town, and village. However, such location information is insufficient for a specific event occurrence area. Therefore, we strive to identify and use more accurate position information estimation. The vocabulary included in the tweet text is analyzed. Therefore, it is not always easy to locate the event, even if it is described. Even if a reference to an event exists, tweets made before going to the event venue or after returning home are not tweets that were actually issued from the event venue. Nevertheless, removing such tweets from the overall data is not easy. One study presented a method to use Flickr data with location information as compensation for shortages of tweets with location information [13]. Some studies have proposed a method of estimating the location from a user's residential area and a method of estimating the location from the IP address of the device [14]. In the position estimation used for this study, the tweet text is analyzed in the same way as in the earlier study. However, it is characterized by estimating more accurate location information around the area where the event occurs, assuming that one already has location information for each municipality. Another feature is that the estimation accuracy is sufficient if it is useful for detecting deflation.

3 TARGET EVENT AND DATA

The target event for this experiment was the visit of the General Public to the Imperial Palace after Accession to the Throne on May 4, 2019. About 141,000 people visited the palace as members of the general public. Their Majesties the Emperor and Empress appeared at the balcony of the Chowa-Den Hall six times to greet visitors who had gathered there. Participants were able to enter from the main gate of the Imperial Palace. The time from 9:30 am to 2:30 pm was the entry time. Tweets with geotags analyzed for this study were sent from areas circumjacent to the Imperial Palace. Those in this range were visitors of the general public who tweeted while they waited or after they left. One can imagine that they would be unable to tweet when moving to Chowa-Den immediately before each appearance, and that they would refrain from tweeting during each appearance. By checking the tweet status, one can estimate the participants' movements: whether waiting or moving.

We chose this event as a target because the people's movements are simple and because it is known exactly when they move. The goal is judging people's movements by analyzing tweets. Therefore, the correct answer must be found to know whether the judgment is correct. It would be interesting to know the movement of people at events such as Halloween parties and the Gion Festival. However, even if Halloween and the Gion Festival were target events of the experiment, the correct answer of people's movements could not be known. People's movements are complex: obtaining accurate information about where and when and how many people are congregating is difficult. Even if the location information were estimated, it cannot be evaluated whether the estimation is correct. By contrast, people's movements are simple during the visit of the General Public to the Imperial Palace after Accession to the Throne. Moreover the movements are known accurately. Evaluation of the experiment results can also be done quantitatively.

The tweets to be analyzed were tweets including geotag “coordinates”. The geotag “coordinate” data represent single latitude and longitude coordinates for the location at which a terminal is used for posting a tweet. Location information is obtainable from the GPS function of the terminal that posts the tweet or the access point connected to the network. It is attached to the tweet it posts. Therefore, the data are highly useful as positioning data. Tweets were extracted during 00:00:00 – 23:59:59 on May 4, 2019: the day the target event took place. The extraction range of tweets, the area, is defined by the following four latitude and longitude coordinates shown in ①–④ in Fig. 1. This is the venue for the target event and the area surrounding the front yard of the Chowa-Den where people moved.

[35.677002, 139.753658]
[35.689604, 139.753658]
[35.689604, 139.761212]
[35.677002, 139.761212]

By filtering the data using the conditions shown above, 198 tweets were extracted. Multiple tweets were posted from the same account. There were 116 unique accounts.

Their Majesties the Emperor and Empress presented their First Appearance at 10 am. Appearances were held six times until 3 pm. People move to the Chowa-Den front courtyard about 20 min before appearances, but they waited until then. Twenty minutes before the appearance, people move and waited for their appearance to hear the words of His Majesty the Emperor. When an appearance is concluded, people will disperse from the front of the Chowa-Den. People can not afford to tweet during this time. In Table 1, it is expressed as "Move". People wait about 30 min for the next appearance to start. Therefore, as Table 1 shows, we estimated how long people would have moved (or stayed).

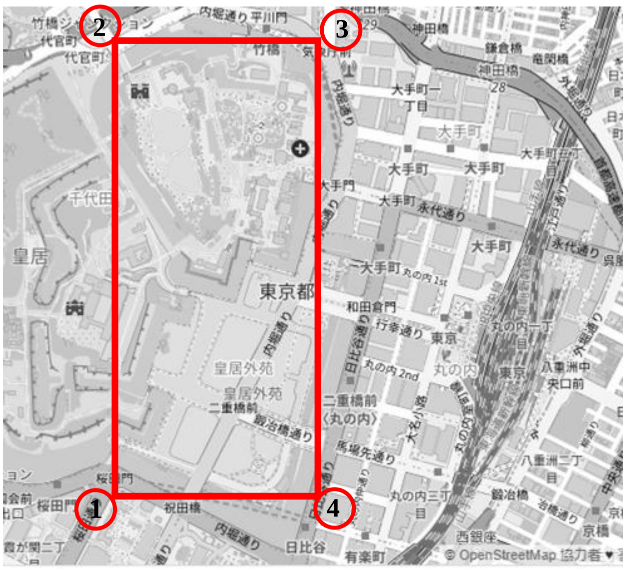

Figure 1: Target Area(in front of the Imperial Palace).

Table 1: Correct answer to estimate.

	time	action of people	
before open gate	– 9:40	Stay	
around appearance	9:40 – 10:10	Move	Deflation
	10:10 – 10:40	Stay	
	:	:	
	14:10 – 14:40	Stay	
around appearance	14:40 – 15:10	Move	Deflation
	15:10 – 16:00	Stay	
event end	16:00 –	Move	Deflation

4 DETERMINATION METHOD

The following two methods are used as the deflation determination method. One is based on the method reported by Endo et al. They succeeded in determining changes in people’s posting on a daily basis, such as cherry blossom viewing time estimation. In our study, a person’s movement is judged in units of minutes instead of days. The other method is derived from real-time burst determination as reported by Ebina et al. Our criteria are reversed to assess a deflation rather than a burst.

4.1 Method Based on the Endo et al. Method

This method uses the tweet posting frequency, which is related closely to the tweet posting interval. Tweet post intervals have high real-time characteristics because they depend on each tweet. However, a certain posted interval must be found to calculate the tweet posting frequency. Usually, since the posting interval for calculating the posting frequency is set larger than the individual tweet posting interval, the real-time property of the analysis using the tweet posting frequency is lower than the analysis using the posting interval.

The method presented by Endo et al. uses a moving average of the frequency of posting tweets to estimate the full bloom of cherry blossoms or other phenomenon. The method calculates the frequency daily and examines differences between the 5-day moving average and the 7-day moving average.

The judgment criterion for the best time is when the 5-day moving average becomes greater than the 7-day moving average and becomes larger than the average of the prior year. In this study as well, a comparative experiment was conducted using this condition. However, the tweet posting frequency is not calculated on a daily basis in this study, but on a 5-min basis. Unlike efforts to infer the best time to view cherry blossoms and so on, we wanted to assess the movement and congestion of crowds of people. Therefore, shorter intervals of 5 min were used instead of 1 day in the frequency calculation.

Table 2: Evaluation using the method of Endo et al..

	Precision	Recall	F-value
3(15-minute) moving average / 5(25-minute) moving average	47.06%	5.76%	10.26%
5(25-minute) moving average / 7(35-minute) moving average	64.81%	25.18%	36.27%
3(15-minute) moving average / 10(50-minute) moving average	48.94%	16.55%	24.73%

From the change in the moving average of the tweet posting frequency, one can examine how accurately the movement of the person in Table 1 can be found. Because it is a moving average, the frequency of posting tweets does not change drastically. One can use two moving averages of different lengths. For people who are actually moving, the short moving average is smaller than the long moving average. We also changed the length of the average of the moving averages to be compared. Table 2 presents results of the quantitative evaluation. Not only was judgment based on the difference between 5 moving average and 7 moving average evaluated; we also evaluated judgment results obtained when using the difference between the 3 moving average and 5 moving average and judgment results obtained when using the difference between the 3 moving average and 10 moving average. The precision is high, but the recall is low. Moving averages are used. Therefore, it is not possible to respond sensitively to changes. Moreover, there are many oversights. Both of those shortcomings engender poor recall.

4.2 Method Based on the Ebina et al. Method

The method of Ebina et al. uses the tweet posting interval instead of the tweet posting frequency for real-time determination. Similarly to assessment of the change of the moving average, burst judgment is applied by the change of multiple tweet posting intervals. Specifically, if the moving average of the tweet interval becomes short, it is judged as a burst. Using this method, deflation is inferred by reversing the judgment conditions. In other words, the deflation occurrence condition is attained when the tweet posting interval becomes longer than before.

Table 3: Quantitative evaluation according to Ebina et al.

	Precision	Recall	F-value
5 number analysis	36.46%	55.56%	44.03%
10 number analysis	42.98%	77.78%	55.37%
15 number analysis	39.51%	50.79%	44.44%

Similar to the discussion presented in the preceding section, we examined how accurately the deflation of the person in Table 1 can be judged under the deflation judgment condition explained above. Table 3 presents quantitative evaluation results. The change in the tweet posting interval is compared with the average of the earlier tweet posting intervals. Each row in the table is compared with the average posting interval of the 5 prior, the average posting interval up to 10 prior, and the average posting interval up to 15 prior. The recall rate of the method of Ebina et al. is high because it reacts in real time. However, its precision is not as good as that achieved when using the method reported by Endo et al. Because it has excellent real-time performance, however, an excessive reaction sensitively causes a decrease in the precision rate. It is possible to judge without overlooking the phenomenon that deflation and the recall and F-value are high.

5 LOCATION ESTIMATION

To improve the accuracy of deflation judgment, we will strive to increase the size and the quality of the dataset used. In the preceding chapter, we estimated people's movements using tweets with geotag "coordinates" that can provide accurate location information. However, only 116 accounts posted the tweets used in that experiment. The number of visitors involved in the general visit was 141,130. Even if the percentage of users who tweet is low, one can infer that tweets are actually posted from more accounts because the data are limited to those with geotag "coordinates" that can specify the position. Therefore, we will perform verification by increasing the number of analysis targets using tweets with unclear positioning. We will use machine learning to extract tweets posted at the target event venue from the group of tweets including "place" that represents Chiyoda ward. We estimate the posting position. The number of tweets to be analyzed is therefore increased. Then the accuracy of deflation determination is evaluated quantitatively. First, we discuss extraction of tweets posted during the visit of the General Public to the Imperial Palace from tweets that are clearly posted in Chiyoda ward.

5.1 ”Place” Data

As a tweet with a geotag for which position information is ambiguous, a tweet for which the geotag metadata is "place_type:city" is used. The geotag "place" has a rectangular range of information represented by four latitudes and longitudes. This rectangle delimits location information at the city level. Compared to the number of tweets that have the geotag "coordinates" as accurate location information, the number of tweets that include only the geotag "place" is larger. It depends on the location, but a difference of about four times in Japan exists overall. The data to be classified by machine learning are tweets with no geotag "coordinates" added on May 4, 2019. Only "place" data representing Chiyoda ward, Tokyo are used. The Imperial Palace is located there. There were 3132 tweets.

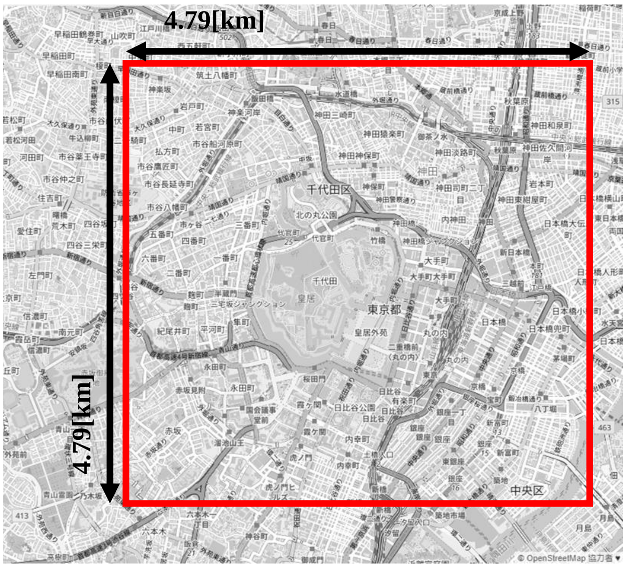

Figure 2: “place” of Chiyoda Ward.

The attached “place” data were confirmed for the 198 tweets with “coordinates” used in the preceding chapter. Results show that 198 "place" data were all the same value. The following four points were recorded.

[35.6686, 139.73]
[35.7052, 139.73]
[35.7052, 139.783]
[35.6686, 139.783]

Figure 2 portrays these four points. The range for this place is the range entirely circumscribing Chiyoda ward.

Tweets that have only the geotag "place" and no geotag "coordinate" are shown in the range of latitudes and longitudes of the four points. They can be narrowed down to the municipality, but the exact tweeted position cannot be found. Therefore, we analyze the tweet contents using natural language processing and consider a method to estimate the user's position more accurately based on the tweet contents. To infer the location, 3132 tweets with the geotag "place" representing Chiyoda ward, where the Imperial Palace is located, are binary-classified using machine learning to estimate whether or not the tweet is from a user who has visited the Imperial Palace. When classifying tweets with only the geotag "place" added by machine learning, teacher data are extracted from text data of tweets with geotag "coordinates" added. For classification, the target data used for estimating the position were text data of the tweet to which only the geotag "place" representing Chiyoda ward was added. We used only tweets posted on May 4, 2019, when the target event was held.

5.2 Vectorization and dimension reduction

To use machine learning, we create teaching data consisting of a set of tweets as a model. These teaching data include text data of tweets with geotag "coordinates". All text data of model tweets are collected to compile a word dictionary. The word dictionary comprises noun words obtained from all words that appear by analyzing the tweet set morphologically and by dividing it into words. This time, we extracted only nouns. Then, assuming that tweets with the geotag “coordinates” are few, we particularly examine nouns for each tweet. MeCab is used to extract the morphemes. Furthermore, by normalization, character strings including only numbers, katakana, and alphabet characters are excluded as stop words. For dictionary data used in MeCab, in addition to the IPA dictionary provided as standard in the morphological analyzer, a user dictionary created from keyword files of "Wikipedia" and "Hatena Keyword" is also used so that minor nouns can be supported.

To convert text data into numerical data that can be processed using machine learning, the data must be vectorized. Bag of Words (BoW) is used for vectorization. With BoW, after word appearances are counted for each tweet, a matrix is generated from the counted words.

The machine learning result is affected by vectorizing the extracted words using morphological analysis. A great amount of noise is included if one simply vectorizes words without consideration of their importance and meaning. Two preprocesses, TF-IDF and LSI, were applied to use the feature vector extracted from each tweet’s contents as optimum data for use in machine learning.

The vectorized features are weighted by TF-IDF, which is a method to weight the words when classifying individual tweets when the number of occurrences of highly important words is high in a tweet set. Term frequency (TF) represents the number of times a word appears in a tweet. Document frequency (DF) represents the number of tweets in which a word appears. In addition, IDF is the logarithm of the reciprocal of DF.

Latent semantic indexing (LSI) is a dimensional compression method using singular value decomposition (SVD). The LSI method specifically examines the latent meaning of words for mitigating over-learning and for reducing learning costs. Indexing synonyms and making synonyms into a vector can be done by indexing the latent meanings of words.

Regarding the number of dimensions to be compressed, the greater the number of dimensions used for machine learning becomes, the higher the calculation cost becomes. Moreover, the processing time increases. As described in this paper, we reduce the dimensions of feature vectors weighted by TF-IDF to 100 dimensions by LSI.

5.3 Incorrect answer area

To estimate the place from which a tweet with only a geotag "place" was posted, we used SVM: a learning model that can execute binary classification by machine learning. The place is classified by SVM from the tweet contents. Actually, SVM has good compatibility with binary classification and high generalization performance. The following ranges for extracting correct and incorrect data are both included in the "place" range (Fig. 2) for Chiyoda ward.

Among the teaching data, the 198 tweets used in the experiment in the preceding chapter were used as correct answers. The tweet group in the range (including the Imperial Palace) where the event occurred is used as correct answer data.

Incorrect answer data were extracted from tweets originating from areas where the event did not occur: not including the Imperial Palace. These tweets have a "coordinates" geotag. The range of the incorrect answer data is the same area

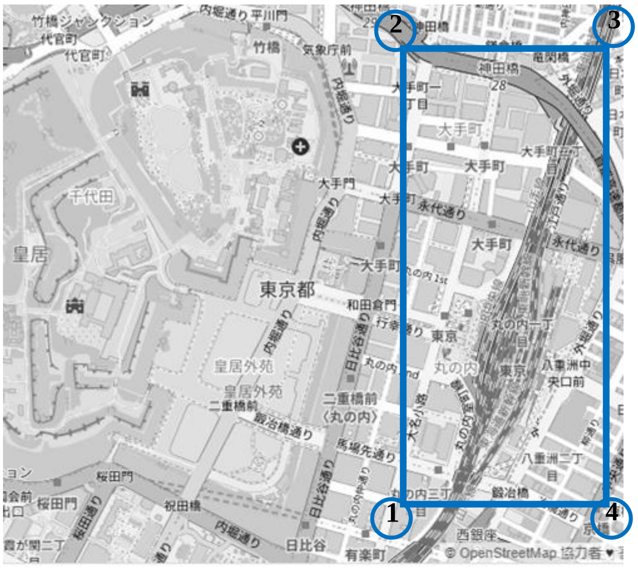

Figure 3: Area of incorrect data.

as that of the correct answer data. This is the target event location: 200 m east of the Imperial Palace. The range of latitude and longitude from the southwest is given below. Figure 3 portrays the approximate range of the extracted incorrect data (not correct data) on the map.

$$[35.677002, 139.763425],$$
$$[35.689604, 139.763425],$$
$$[35.689604, 139.770979],$$
$$[35.677002, 139.770979]$$

By extracting tweets to which "coordinates" were added under the conditions depicted in Fig. 3, 866 tweets were eventually extracted. Then the text set with correct and incorrect data combined was used as a learning model.

5.4 SVM

We used a classification method with Support Vector Machine (SVM), which is capable of binary classification positioning of a person who was at the scene of a specific event or incident from a message to which only the geotag "place" was added. Actually, SVM has been adopted in many earlier studies as a method to estimate user attributes and position from tweet contents. For the SVM kernel, we adopt a linear kernel that is often used when classifying large-scale data, sparse data, and text data. In addition, cost parameter C is set to the default value of 1 for learning.

The linear kernel SVM is a classifier that constructs a hyperplane that maximizes the shortest distance (margin) between the classification boundary and the training data.

For machine learning model evaluation, cross validation is undertaken by dividing the data into an arbitrary number K using the Stratified K-Fold method. The tweet structure of the evaluated model is the following. It consists of data of two types. One data type comprises 198 tweets with geotag "coordinates" extracted in the correct answer range (Fig. 1), which is regarded as having caused people to loiter and flow because of the occurrence of events. The other data type comprises 866 tweets with geotag “coordinates” extracted in the range (Fig. 3) where the target event (Visit of the General Public to the Imperial Palace after Accession to the Throne) has not occurred. The 1064 tweet data, which include data of these two types, are divided into five portions: 4/5 are training data; 1/5 are test data. The tweet vector generated from the training data is the explanatory variable. However, with classification by SVM, for the objective variable, a binary label assigned to the test data is 1 for a correct answer and -1 for an incorrect answer. The training data and test data are exchanged. Classification is performed using SVM five times in all. The average value of the results of five cross-validations was used for estimating the model accuracy. The machine learning library scikit-learn was used to implement SVM. This time, K=5 split cross validation was performed. The average value of the classification correct answer rate for five times was calculated. The classification accuracy rate is an index showing how well the classifier can classify. The classification accuracy rate can be expressed as equation (1).

$$\text{Classification correct answer rate} = \frac{\text{Number of successful classifications}}{\text{Number of evaluation cases}} \tag{1}$$

As a result, it was 80.64%.

5.5 Estimated result

The classification target is 3132 tweets, with only the geotag “place” that represents Chiyoda ward. The SVM has assigned a label of 1 to tweets that are judged to be in the correct answer range, and a label of -1 to tweets that are judged to be in the incorrect answer range. From extraction of the tweets with correct labels, 1750 tweets were output as correct answers.

These 1750 tweets are combined with data of 198 correct tweets with the geotag "coordinates". Using 1948 tweets, we conduct analysis using the Endo et al. method and the Ebina et al. method. By calculating the precision, recall, and F-value, the accuracy of a deflation judgment can be evaluated quantitatively when analysis is performed by adding tweets estimated as having been posted in the range presented in Fig. 1 by machine learning.

Similarly to Table 2, Table 4 presents results of quantitative evaluation using the method described by Endo et al. As in Table 2, parameters of three types are shown. In Table 2, we conducted evaluation using only 198 tweets having the geotag “coordinate”. However, in Table 4, as a result of position estimation, the number of tweets used for evaluation was 1948, which is larger than those used for Table 2. The number of tweets has increased, but the precision and recall have not improved. Table 5 presents results of quantitative evaluation using the method of Ebina et al. In Table 5, the number of tweets used for evaluation is greater than in Table 3. Compared to the values presented in Table 3, the recall is not very good, but the precision is good.

We were able to increase, by nearly ten-fold, the number of tweets used to judge deflation by machine learning: from

Table 4: Evaluation using the method of Endo et al.

	Precision	Recall	F-value
3(15-minute) moving average / 5(25-minute) moving average	48.89%	15.83%	23.91%
3(15-minute) moving average / 10(50-minute) moving average	43.75%	25.18%	31.96%
5(25-minute) moving average / 7(35-minute) moving average	47.37%	19.42%	27.55%

Table 5: Evaluation using the method of Ebina et al.

	Precision	Recall	F-value
5 number analysis	56.55%	50.53%	53.38%
10 number analysis	57.98%	67.02%	62.17%
15 number analysis	58.49%	53.72%	56.01%

198 to 1948. The tweet posting position was estimated using SVM, but its classification accuracy rate in cross-validation is about 80%; it includes about 20% noise. Although the number of tweets used to judge deflation was increased by about 10 times, noise is included in that result. Therefore, the deflation judgment accuracy might or might not improve.

5.6 Estimated results with changed teaching data

Next, we changed the teaching data. No tweets are used as teaching data other than 198 tweets of correct answer data and 866 tweets of incorrect answer data. These are all tweets with the geotag "coordinate" which records the exact location. All of these were used as described in the preceding section, but the number of correct and incorrect data is unbalanced. Incorrect answer data are about five times as numerous as correct answer data. What we learned from correct data by machine learning is less than what we learned from incorrect data. Therefore, we randomly extracted only 495 tweets without using all incorrect answer data. This is exactly 2.5 times the number of correct data. First, a five-fold cross validation test was performed with 198 tweets as correct answer data and 495 tweets as incorrect answer data. Results show that the classification accuracy rate was 68.84%. Compared to the use of 866 tweets as incorrect answer data, the decomposition correct answer rate decreased by 10% or more.

We also investigated the case in which only 198 tweets were selected randomly from 866 tweets as incorrect answer data. The numbers of correct answer data and incorrect answer data are the same. When five-fold cross validation was applied using these 396 tweets, the decomposition correct answer rate was 90.86%, an improvement of 10% or more.

Table 6: Evaluation using the method of Endo et al. for teaching data of 198 correct data and 495 incorrect data.

	Precision	Recall	F-value
3(15-minute) moving average / 5(25-minute) moving average	43.02%	16.79%	24.15%
3(15-minute) moving average / 10(50-minute) moving average	34.84%	25.46%	29.42%
5(25-minute) moving average / 7(35-minute) moving average	43.49%	21.83%	29.07%

Table 7: Evaluation using the method of Ebina et al. for teaching data of 198 correct data and 495 incorrect data.

	Precision	Recall	F-value
5 number analysis	49.77%	53.58%	51.60%
10 number analysis	46.17%	67.78%	54.93%
15 number analysis	53.69%	60.40%	56.85%

The number of teaching data and the decomposition accuracy rate are not always correlated. Moreover, the decomposition correct answer rate does not correlate with the number of incorrect answer data in the teaching data.

Five-fold cross validation was applied to a total of 693 tweets, comprising 198 tweets as correct answer data and 495 tweets as incorrect answer data. Results show that the decomposition accuracy rate was 68.84%. All of these 693 tweets were learned as teacher data. The position of 3132 tweets for which accurate position information was unknown was estimated. The geotag "coordinate" is not recorded in these 3132 tweets, but the geotag "place" is recorded. They are tweets posted in Chiyoda Ward. From position estimation by SVM, 350 tweets out of 3132 tweets were output as correct answers. It was estimated that these 350 tweets were posted within the target area of Figure 1. These 350 tweets are 548 tweets in addition to the 198 tweets for which the posting position was originally clear. With these 548 tweets, the methods of Endo et al. and of Ebina et al. were used to evaluate deflation, with results presented respectively in Tables 6 and 7.

Only 350 tweets were added by estimating the position. Compared with Table 4 and Table 5, for which 1750 tweets were added by position estimation and then deflation judgment was performed, there is little difference overall. In fact, it is slightly worse. The result is not an improvement compared to Tables 1 and 2, which show evaluation results of deflation judgment performed using only 198 tweets with the correct posting position. This can be regarded as a natural result because the decomposition correct answer rate by five-fold cross validation is also poor and data are few.

Table 8: Evaluation using the method of Endo et al. for teaching data of 198 correct data and 198 incorrect data.

	Precision	Recall	F-value
3(15-minute) moving average / 5(25-minute) moving average	49.24%	15.75%	23.87%
3(15-minute) moving average / 10(50-minute) moving average	42.14%	24.13%	30.69%
5(25-minute) moving average / 7(35-minute) moving average	44.53%	17.61%	25.24%

Table 9: Evaluation using the method of Ebina et al. for a teaching data of 198 correct data and 198 incorrect data.

	Precision	Recall	F-value
5 number analysis	56.95%	50.29%	53.41%
10 number analysis	55.85%	64.22%	59.74%
15 number analysis	54.99%	48.70%	51.66%

For five-fold cross validation applied with 396 tweets with the same number of correct and incorrect data, the classification correct answer rate was 90.86%. Position estimation was performed using only these 396 tweets as teacher data. By machine learning, we obtained data estimated as the correct answer for 1738 tweets from 3132 tweets that have only the geotag "place". In other words, it was estimated that 1738 tweets out of 3132 tweets were posted in the target area presented in Fig. 1. Adding 198 tweets with known location information to these 1738 tweets yields 1936 tweets. Using these 1936 tweets, deflation judgment was performed using the methods of Endo et al. and of Ebina et al. Table 8 and Table 9 respectively present the results.

Compared with Tables 4 and 5 with deflation judgment performed by adding the 1759 tweets obtained using position estimation, it is worse overall. Furthermore, these are not good compared to Tables 6 and 7, where only 350 tweets were added by location estimation. The classification correct answer rate by five-fold cross validation yields the best result, but high accuracy of position estimation for 3132 tweets, which have only the geotag "place", is not guaranteed. The assumption that 1738 out of 3132 tweets were posted in the target area of Fig. 1 is slightly high. Tokyo Station is outside of the target area. Therefore, more tweets should be posted outside the target area.

6 DISCUSSION

Many data must be used to judge deflation in real time. Few tweets have geotag “coordinates” that can specify the position. Therefore, we use machine learning to estimate the locations of tweets that have a geotag “place”, which is ambiguous location information for each municipality. Deflation was inferred from tweets that were presumed to have been posted at a specific location. The method presented by Endo et al. did not improve the judgment accuracy, but the method presented by Ebina et al. did. We have improved the accuracy rate of deflation judgment using the method of Ebina et al. by increasing the number of tweet data that can be used by estimating the location information. Results show that the recall has not improved, but the F-value has improved.

In other words, SVM location estimation increased the number of tweets posted at the target event location. The classification accuracy rate was not bad, as indicated by five-fold cross validation performed only with tweets with the geotag "coordinate". Nevertheless, because the position estimation accuracy is insufficient, when judging the movement of people using the obtained tweets, the judgment accuracy might decrease instead of increasing. Five-fold cross validation showed that tweets posted in the target area presented in Fig. 1 were used as correct answer data; tweets posted in the area depicted in Fig. 3 were used as incorrect answer data. These areas are not adjacent and might have contributed to a better classification accuracy rate. It might be true that tweets posted in the area presented in Fig. 2 excluding the tweets posted in the area presented in Fig. 1 should have been regarded as incorrect data. In that case, if those tweets were used as teaching data for machine learning, the tweets posted near the target area in Fig. 1 would become noise and therefore reduce the classification accuracy rate. This point must also be investigated carefully.

Although SVM was adopted for machine learning, various other methods are available. Future studies must compare results obtained from their use. For this study, the data of the tweet contents learned by machine learning were only nouns. Verbs, because of conjugation variants, were avoided. As one might expect, they are difficult to handle. It is also worth considering how verbs and adjectives should be used, in addition to word-dependence. The use of information such as accounts, rather than the tweet body text, should also be considered. Additionally, it is necessary to consider methods other than judging the movement of people as a method of using tweets for which the position is estimated. Various target events can be considered when judging the movements of people according to similar stimuli. These points will be addressed in future research.

7 CONCLUSION

Numerous data must be accumulated to support real-time judgment of the movement of people, especially for deflation, whereby people disappear and spread out. Few tweets include accurate location information. Therefore, for a given area, we used machine learning to estimate the locations of tweets that have a “place” geotag, which is ambiguous location information. Deflation was inferred from tweets that were presumably posted at a certain place. Because of imperfect estimation of position information, the deflation judgment accuracy did not necessarily improve. However, the deflation judgment accuracy improved in some cases.

Some room exists for improvement of position estimation. Future tasks are consistent and reliable improvement of deflation judgment accuracy.

8 ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 20K12081.

REFERENCES

[1] T. Sakaki, and Y. Matsuo, Twitter as a Social Sensor: Can Social Sensors Exceed Physical Sensors? Journal of Japanese Society for Artificial Intelligence, Vol. 27, No. 1, pp.67-74 (2012).

[2] J. Kleinberg, Bursty and hierarchical structure in streams, In Proc. Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp.91-101(2002).

[3] Y. Zhu, and D.Shasha, Efficient Elastic Burst Detection in Data Streams, Proc. Ninth ACM SIGMOD Inter-national Conference on Knowledge Discovery and Data Mining, ACM, pp.336-345(2003).

[4] D. Shasha, and Y. Zhu, High Performance Discovery in Time Series, Techniques and Case Studies (Monographs in Computer Science), Springer-Verlag(2004).

[5] X. Zhang, and D. Shasha, Better Burst Detection, Proc. 22nd International Conference on Data Engineering, IEEE Computer Society, pp.146-149 (2006).

[6] K. Ohara, K. Saito, M. Kimura, and H. Motoda: “Burst Detection in a Sequence of Tweets based on Information Diffusion Model”, DBSJ journal, Vol.11, No. 2, pp.25-30(2012).

[7] R. Ebina, K. Nakamura, and S. Oyanagi, A Proposal for a Real-Time Burst Detection Method, DBSJ Journal, Vol.9, No.2, pp.1-6(2010).

[8] R. Ebina, K. Nakamura, and S. Oyanagi, A Proposal for a Real-time Burst Analysis Method, IPSJ TOD, Vol.5, No.3, pp.86-96(2012).

[9] M. Endo, Y. Shoji, M. Hirota, S. Ohno and H. Ishikawa, On Best Time Estimation Method for Phenological Observations Using Geotagged Tweets, IWIN2016, pp.205-210 (2016).

[10] M. Endo, M. Hirota, S. Ohno, and H. Ishikawa, Best-Time Estimation Method by Region and Tourist Spot using Information Interpolation, IWIN2016, pp.209-216 (2017).

[11] O. Ozer, R. Heri, and N. Kjetil, Locality-adapted Kernel Densities for Tweet Localization, SIGIR'18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 1149-1152(2018).

[12] T. Morikuni, M. Yoshida, M. Okabe, and K. Umemura, Geo-location Estimation of Tweets with Stop Words Detection, Journal of Information Processing, Transaction on Database Vol.8, No.4, pp.16–26(2015).

[13]T. Mansouri, AZ. Ravasan, and MR Gholamian, A novel hybrid algorithm based on k-means and evolutionary computations for real time clustering, Int J Data Warehous Min (IJDWM) 10(3):1–14 (2014).

[14] J. Eisenstein, B. O’Connor, NA. Smith, and EP. Xing, A latent variable model for geographic lexical variation, In Proceedings of the 2010 conference on empirical methods in natural language processing, Association for Computational Linguistics, pp 1277–1287(2010)

(Received October 31, 2020)
(Accepted March 30, 2021)

Takuma Toyoshima is currently working as vocational training instructor in a course of training that production support with ICT at the Polytechnic center Shiga. He received a Master of degree in Industrial Engineering from the NIAD-QE Tokyo in 2020.

Takuo Kikuchi received the Ph.D degree in engineering from Kyushu Institute of Technology. He is currently a professor, dept. of electrical and computer engineering at Polytechnic University. His research interests include “Information network cabling” and “Skill analysis and evaluations”. He is a member of the JSEE, IEICE and JSET Japan.

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.

Shigeyoshi Ohno earned a M.Sci. and a 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.