Estimating deflation representing people spreading in stream data and estimating a specific position

Takuma Toyoshima

Polytechnic Center Shiga,
Ohtsu-shi, Shiga, Japan
Email: m18303@uitec.ac.jp

Masaki Endo, Takuo Kikuchi and Shigeyoshi Ohno*

Division of Core Manufacturing,
Polytechnic University,
Kodaira-shi, Tokyo, Japan
Email: endou@uitec.ac.jp
Email: kikuchi@uitec.ac.jp
Email: ohno@uitec.ac.jp
*Corresponding author

Hiroshi Ishikawa

Graduate School of System Design,
Tokyo Metropolitan University,
Hino-shi, Tokyo, Japan
Email: ishikawa-hirosi@tmu.ac.jp

Abstract: With the expanded use of social media such as Twitter in recent years, it has become easy to add various information such as location data using mobile devices. Using those data, one can observe the real world without using physical sensors. Therefore, social media have high operational value as social sensors. As described herein, we aim to support decision-making for people who intend to visit a specific place at which an event or some trouble recently occurred. After proposing a method of real-time extraction of data reflecting a burst state showing people's concentration, their inactivity, and continuous flow and dispersion, we confirm the method's effectiveness. We will also try to estimate the location information of tweets for the purpose of further improving the estimation accuracy. Since few tweets have accurate location information, we use the content text of the tweet to find the tweet posted at the event occurrence location by machine learning. We will study changes in the accuracy of the proposed method due to the increase in the data to be analysed.

Keywords: deflation; burst; Twitter; real-time extraction; social sensor.

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

Masaki Endo earned his BE from the Polytechnic University, Tokyo and graduated from the course of Electrical Engineering and Computer Science, Graduate School of Engineering Polytechnic University. He received an ME from the NIAD-UE, Tokyo. He earned his PhD in Engineering from the 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.

Takuo Kikuchi received his PhD in Engineering from the Kyushu Institute of Technology. He is currently a Professor in the Department 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.

Shigeyoshi Ohno earned his MSc and DrSci from the Kanazawa University, and DrEng from the 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 BS and PhD in Information Science from the University of Tokyo. After working for Fujitsu laboratories and becoming a Full Professor at the Shizuoka University, he became a Full Professor at the 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.

1 Introduction

In modern society, with computerised devices and systems, many means exist to acquire information of various types in real-time. One means is Twitter, a microblogging service that shares short sentences called 'tweets' of 140 or fewer characters. It is widely used throughout the world, as it is in Japan. Many users regard it as a medium by which they can post recent information casually. Because posting of location information can be done with so-called geotags easily via a smartphone, it is a social medium that can immediately notify many people of what is happening and where. Based on these characteristics, it is anticipated for use as a social sensor for observing the real world without using high-cost physical sensors (Sakaki and Matsuo, 2012).

Social sensors can reveal a situation in real-time even if one is not present on the scene. For example, if one can guess the best time for cherry blossoms and autumn leaves before visiting a place, one could actually go there without a worry that cherry blossoms have not yet bloomed or that they have already fallen. If public transportation is halted and one knows that people are crowding into stations, then one can avoid those crowds. If a person knows that the congestion has eased, then the person could stop avoiding the area. If a reveller 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 has settled down, it would not be so interesting. Alternatively, to avoid a raucous party, one might wait until after it has settled. Social sensors would be useful to inform people to make such choices.

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. It does not make sense to analyse today's event using yesterday's data. Assuming that one is not actually present in a certain place, then Twitter data can be used to avoid deployment of expensive sensors to ascertain the concentration of people in real-time. This study was conducted to produce a means of real-time estimation of human motion by analysing geotagged tweets.

2   Related work

Kleinberg (2002) proposes a method for modelling text stream bursts and for extracting structures. This method is based on modelling a stream using an infinite state automaton. A salient benefit of Kleinberg's approach is that it can represent 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 Shaha (2003), Shasha and Zhu (2004) and by Zhang and Shasha (2006) examine bursting algorithms that monitor bursts efficiently over multiple window sizes. These techniques enable near 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.

Ebina et al. (2010, 2012) propose a method for real-time burst detection, which is achieved by inference of whether or not each event (each tweet posting) is a burst. 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, it remains unclear whether the burst state continues or immediately ends solely by the burst occurrence.

Endo et al. (2016, 2017) use the moving average to make full-fledged decisions. 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 quantised by the window size. Using the tweet posting frequency requires setting of a certain time interval for frequency calculation. This fixed interval impairs real-time performance.

This study examines a proposed method to detect convergence of the burst state as well as real-time properties and burst generation. In reference to the method described by Ebina et al., this research ensures real-time capability using the tweet posting interval

instead of the tweet posting frequency in a fixed period. In addition, this research was conducted to detect when posts decrease and return to a normal state, not only in the burst state where tweets post increase rapidly, by combining moving averages similarly to the method explained by Endo et al.

Large amounts of tweet data are necessary to estimate people's movements in real-time. However, few tweets include any location information. Furthermore, much location information is ambiguous. Moreover, tweets with accurate location information are even fewer. Therefore, research is being conducted to obtain location information from tweet contents (Ozer et al., 2018; Morikuni et al., 2015). Methods have been designed to obtain location information by analysing 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 location of the object. However, for the present study, it is assumed that you have location information for each city, ward, town, and village. Such location information is not sufficient for a specific event occurrence area. So we strive to identify and use more accurate location information estimation. Since the vocabulary contained in the tweet text is analysed, it is not always easy to locate the event if it is mentioned. 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 issued from the event venue. Nevertheless, removing such tweets from the overall data is not easy.

3 Our proposed method

To detect deflation after convergence of an event, one must capture the phenomena of people concentrating and spreading at a specific location in real-time. The real-time performance can be achieved by monitoring each tweet posting time and its interval. Then, using Fisher's exact test, the deflation criteria are defined.

3.1 Examination of real-time analysis methods

Actual tweet information is used to verify the proposed method. Because it uses collected data, this is not a real-time experiment. However, real-time determination can be achieved because the proposed method uses no information after the determination time.

To take advantage of real-time capabilities, the tweet frequency is not used in our proposed method. To determine the frequency, a fixed time interval must be determined. This finite time interval impairs real-time performance. Specifically, the tweet time interval is used instead of the frequency with which tweets are posted. Letting $c(h, t)$ be the time interval of the tweets posted at time t, then parameter h represents the time interval from previously posted tweet h. The $c(h, t)$ divided by h is expressed as $avg(c(h, t))$. It is the average value of the tweet occurrence interval from the tweet that occurred h times before time t. If tweets increase, then $avg(c(h, t))$ becomes smaller. If the tweet is lost, it takes a large value.

The timestamp at which the tweet is posted is expressed in seconds. If there are many tweets, then the timestamps of multiple tweets will be the same. For example, the time points are 2 s and 7 s in Figure 1. Each four tweets and two tweets are posted at the same time. In this case, the value of the function $c(h, t)$ might be zero or not determined. However, it was not really posted at the same time, four tweets were posted between

$1.5 \leq t < 2.5$. Therefore, as shown in the lower part of Figure 1, the value of function $c(h, t)$ is made to be determined uniquely by assuming that the tweet is posted by dividing one second. For Figure 1, the $c(h, t)$ values are shown in Table 1.

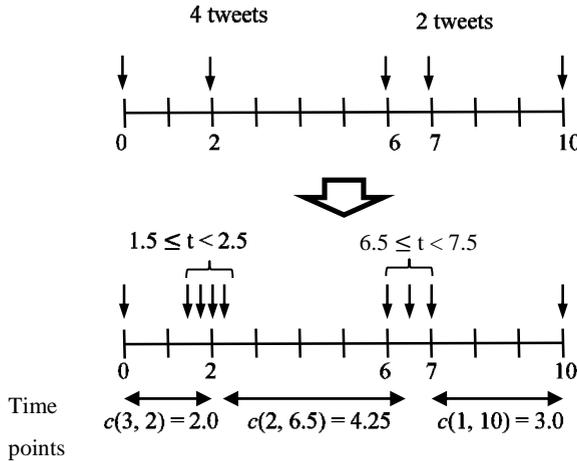**Figure 1**   Example of generating function $c(h, t)$

Table 1   Example of the values of function $c(h, t)$

h	$c(h, 2)$	$c(h, 2.25)$	$c(h, 6)$	$c(h, 6.5)$	$c(h, 7)$	$c(h, 10)$
1	0.25	0.25	3.75	0.5	0.5	3.0
2	0.5	0.5	4.0	4.25	1.0	3.5
3	2.0	0.75	4.25	4.5	4.75	4.0
4		2.25	4.5	4.75	5.0	7.75
5			6	5.0	5.25	8.0

In the method presented by Ebina et al., the magnitude of this value is compared to estimate whether the crowded state is crowded (tweets posting increase) or not. The parameter is used to adjust the comparison. If it is not crowded, then the method will over-react to a slight change just as a tweet is accidentally posted: it will be judged falsely that a burst has occurred. Noise suppression can be achieved using this parameter to reduce the sensitivity.

3.2   *Fisher's exact test idea*

Fisher's exact test is a statistical test that determines the bias of data divided into two categories.

As an example, one can consider a case in which the efficacy is verified in a clinical trial of a new medicine. The verification is performed of data obtained by dividing the patients into four according to two exclusive criteria of *A* and *B*.

A Number of patients receiving new medicine.

$\bar{A}$ Number of patients receiving placebo (no-medicinal).

B Number of patients whose condition recovered thereafter.

$\bar{B}$ Number of patients whose condition did not recover.

When data are divided in this manner, verification is possible using a contingency table such as that shown in Table 2.

Table 2 2×2 contingency table

|	*B*	$\bar{B}$
A	*a*	*b*
$\bar{A}$	*c*	*d*

Using a contingency table as shown in Table 2, an index s representing how the disease condition has been improved is definable as shown below.

$$s = ad - bc \tag{1}$$

If the value of *s* is positive, then the new medicine has a 'positive' effect on improvement of the medical condition. It has a medicinal effect.

Fisher's exact test is very simple and can be applied if it can be dropped into a contingency table. Using this concept, analysis particularly addressing individual tweet intervals is performed. Specifically, we propose a method to divide *n* tweets into four regions and analyse the number of tweets in four regions. Figure 2 shows that the four areas are generated by setting threshold values on the vertical and horizontal axes. Blue dots in Figure 2 represent individual tweets. The threshold value on the vertical axis is the average value of *n* tweets. The horizontal axis divides *n* tweets arranged in time series into two equal halves.

Figure 2   Tweet segmentation diagram using the proposed method (see online version for colours)
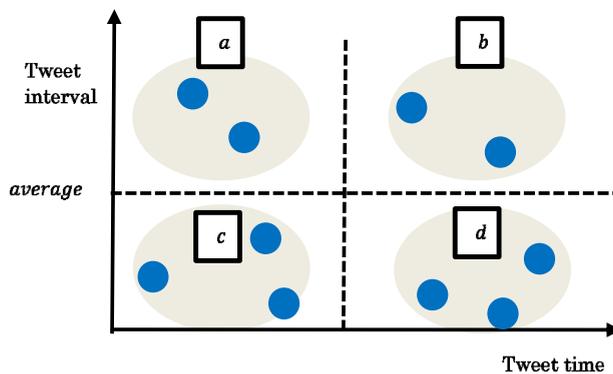
The values of *a*, *b*, *c*, and *d* in the contingency table when doing analyses with *n* pieces are defined as shown below.

a   The first half of the extracted tweets group and the tweet interval is greater than the threshold (average value).

b   The second half of the extracted tweet group and the tweet interval is greater than or equal to the threshold.

c   The first half of the extracted tweets group and the tweet interval is smaller than the threshold.

d   The second half of the extracted tweet group and the tweet interval is smaller than the threshold.

The threshold on the vertical axis representing the tweet interval is the average value of n tweet intervals. The average value can be the arithmetic mean value, the geometric mean value, the median value, or the average value of the maximum value of the minimum value instead of the average value of all values. In the proposed method, the most common arithmetic mean value was used as the threshold value.

When the number of tweets in region a and region d increases and $s > 0$, the tweet interval is narrowed. Thereby, people are concentrated and the value becomes inactive: it is a burst. However, when the number of tweets in region b and region c is increased and $s < 0$, the tweet interval is widened such that people start to spread and move, showing deflation. When $s = 0$, no change occurs in the state. Deflation is estimated in this study using $s < 0$ as the deflation criterion.

This method does not require a dataset for all time period. It is possible to analyse in real-time with only n tweets.

4   Comparative experiments

We conducted an experiment using actual tweet information to verify the proposed method. It is not a real-time experiment because it uses collected data. However, because the proposed method uses no information after the determination time, it can perform real-time determination.

Experiments are also being conducted for comparison using the tweet posting frequency. Those comparison experiments lack real-time capability.

4.1   Data collection

The target event for this experiment was the visit of the General Public to the Imperial Palace after accession to the throne on 4 May 2019 and to the Palace for the New Year's Greeting on 2 January 2020. About 141,000 and 69,000 people visited as members of the general public. Their majesties the emperor and empress appeared at the balcony of the Chowa-Den Hall to greet the visitors who gathered there six times (in 2019) and five times (in 2020). Six appearances were held on an hourly schedule starting at 10 am in 2019. Five appearances were held on an hourly schedule starting at 11 am in 2020. Participants were able to enter from the main gate of the Imperial Palace. The time from 9:30 (10:30 in 2020) am to 2:30 pm was the entry time. Analysed tweets with geotags were made in the area around the Imperial Palace. Those in this range were visitors of the general public who tweeted. They tweeted while they were waiting or after they left. We 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.

Figure 3 Target area (see online version for colours)
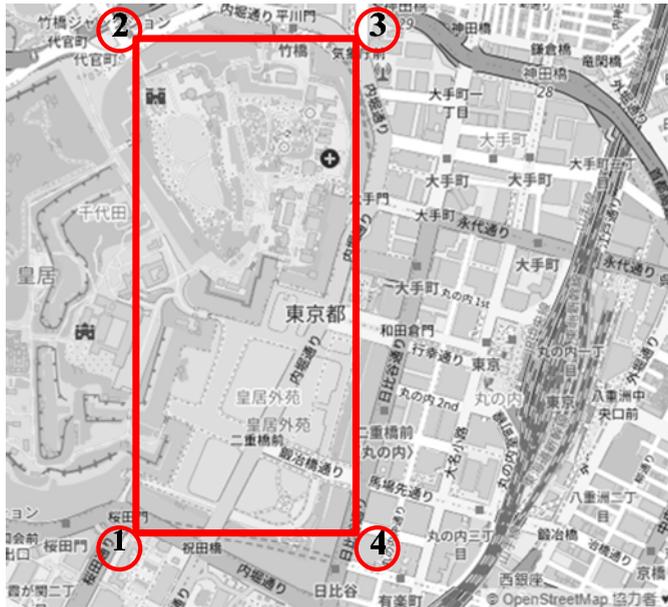
We chose this event as a target because the people's movements are simple and it is known exactly when they move. The goal is judging people's movements by analysing tweets. Therefore, we must find the correct answer to know whether the judgment is correct. It would be interesting to know the movement of people at events such as Halloween 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 is not 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 is estimated, it cannot be evaluated whether the estimation is correct. People's movements are simple in 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 can also be done quantitatively.

The tweets to be analysed were tweets including geotag 'coordinates'. The geotag 'coordinates' data represent the location at which the terminal used for posting a tweet is represented as a single latitude and longitude. Therefore, it is highly useful as location data. The extraction periods of the tweets was 00:00:00–23:59:59 on 4 May 2019 and on 2 January 2020. The extraction range of tweets was determined empirically for the following four latitude and longitude ranges.

$$[35.677002, 139.753658]\,[35.689604, 139.753658]$$
$$[35.689604, 139.761212]\,[35.677002, 139.761212]$$

By filtering under the conditions shown above, 327 tweets (including 198 tweets posted on 4th May) were extracted. Of these, 216 accounts were tweeted.

Table 3   The correct answer to estimate (see online version for colours)

	Time	*Action of people*	
Before open gate	–9:40	Stay	4th May
	–10:40		2nd January
Around appearance	9:40–10:10	Move	Deflation
	10:10–10:40	Stay	4th May
Around appearance	10:40–11:10	Move	Deflation
	⋮	⋮	
Around appearance	13:40–14:10	Move	Deflation
	14:10–14:40	Stay	
Around appearance	14:40–15:10	Move	Deflation
	15:10–16:00	Stay	
Event end	16:00–	Converge	Deflation

Their majesties the emperor and empress presented their first appearance at 10 am (or 11 am). Appearances were held six times until 3 pm. People move to the Chowa-Den front courtyard about 20 min before appearances, but they wait until then. When an appearance is concluded, people will disperse from the front of the Chowa-Den. People wait about 30 min for the next appearance to start. Therefore, as Table 3 shows, we estimated how long people would have moved (or stayed).

4.2   Comparative experiment results: Endo et al.

This method uses the tweet posting frequency in a reciprocal relation with the tweet posting interval. Tweet post intervals have high real-time characteristics because they depended on each tweet. However, a certain time interval must be found to calculate the tweet posting frequency. Usually, the real-time performance is low because the time interval is larger than the tweet posting interval.

The method reported 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 five-day moving average and the seven-day moving average.

The judgment criterion for the best time is when the five-day moving average becomes greater than the seven-day moving average and becomes larger than the average of the prior year. A comparative experiment was conducted using this condition. However, the tweet posting frequency is not calculated on a daily basis, but on a five-minute basis. Unlike the best time to see cherry blossoms and so on, we wanted to know the movement and crowdedness of people. Therefore, a short time interval was used.

The tweet frequency in five minutes at time t is expressed as $e(1, t)$: $e(j, t)$ represents the number of tweets posted in $(5 * j)$ minutes from time $(t – 5 * j)$ to time t. The average of $e(j, t)$ divided by j is represented by $avg(e(j, t))$. Function e is used to set the burst estimation condition as

$$ave\left(e(a,t)\right) > ave\left(e(b,t)\right)$$
$$\text{and } ave\left(e(a,t)\right) > ave\left(e(\text{one day},t)\right).$$
(2)

As the inverse condition of the burst estimation condition, we set the convergence condition of deflation as

$$ave\left(e(a,t)\right) < ave\left(e(b,t)\right)$$
$$\text{and } ave\left(e(b,t)\right) < ave\left(e(\text{one day},t)\right).$$
(3)

Figure 4 represents the judgment under the burst convergence condition.

Figure 4   Deflation determination using 35-minute (orange line) and 25-minute moving average (blue line) (see online version for colours)
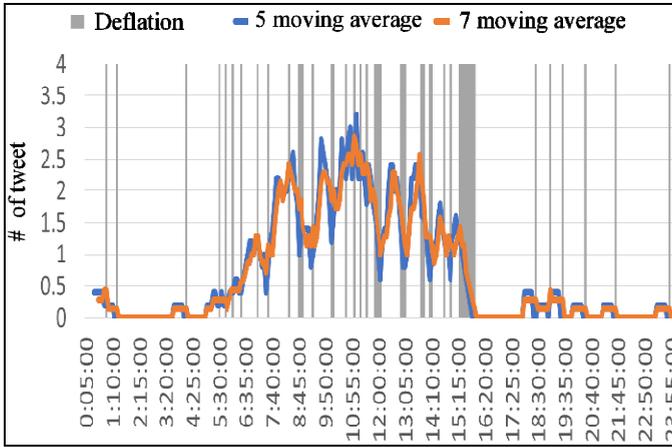

Table 4   Quantitative evaluation using the Endo method (see online version for colours)

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

Similar to Endo's five-day moving average and seven-day moving average, because $a = 5$ and $b = 7$, the 25-minute moving average and 35-minute moving average will be used. The grey parts in the figure show the time periods satisfying the judgment conditions.

From Figure 4, the inactivity disappeared: spreading of people was observed at the time after 15:00 of the last appearance. However, both estimation results from equation (3) include many erroneous estimation results. To elucidate changes at intervals shorter than five minutes, a correct judgment can probably be made by changing to a means having more real-time capability.

Next, we evaluate the deflation judgment result quantitatively. As the evaluation index, the precision, recall, and F-value, which are commonly used in information extraction, were used for comparison. Regarding the parameters, the three sets of $a = 3$, $b = 5$, and $a = 5$, $b = 7$, and $a = 3$, $b = 10$ are applied to equation (3). Then results of

deflation judgment are compared. The judgment accuracy is apparently best with parameters $a = 3$ and $b = 10$.

The recall rate was lower than the precision rate, probably because detection was not possible many times, although deflation actually occurred. This result can be confirmed from the time period during which the polygonal line is decreasing, as shown in Figure 4, and from the delayed timing of the deflation determination.

4.3   Comparative experiment results: Ebina et al.

The comparative experiment results obtained using the method presented by Ebina et al. (2012), are shown in Figure 5. The horizontal axis represents time points. The vertical axis represents tweet occurrence intervals. The blue line graph in the figure is the average tweet posting interval (average of five tweets), representing $avg(c(5, t))$. The left of the figure is the early morning. Therefore, the tweet occurrence interval becomes large. A line is also apparent on the upper left. The right side of the figure is midnight. As the tweet interval increases, there is also a blue line on the upper right.

The crowded (burst) judgment conditions according to the method reported by Ebina et al. are the following.

$$ave(c(1, t)) < 0.4 * ave(c(5, \ t-1))$$
$$\text{or } ave(c(2, t)) < 0.4 * ave(c(5, t-2))$$
$$\text{or } ave(c(3, t)) < 0.4 * ave(c(5, t-3)) \tag{4}$$
$$\text{or } ave(c(4, t)) < 0.4 * ave(c(5, t-4))$$
$$\text{or } ave(c(5, t)) < 0.4 * ave(c(5, t-5)).$$

Reverse the burst judgment conditions and set the deflation judgment conditions as follows.

$$ave(c(1, t)) > 0.4 * ave(c(5, t-1))$$
$$\text{and } ave(c(2, t)) > 0.4 * ave(c(5, t-2))$$
$$\text{and } ave(c(3, t)) > 0.4 * ave(c(5, t-3)) \tag{5}$$
$$\text{and } ave(c(4, t)) > 0.4 * ave(c(5, t-4))$$
$$\text{and } ave(c(5, t)) > 0.4 * ave(c(5, t-5)).$$

The deflation judgment is apparently excessive because the interval between the specific tweet and the tweet posted immediately before it is longer than before. It is designated as a burst. There are times when the appearance time period and time when the event ended. People who were inactive started moving and were able to judge correctly, but reading the spread of people correctly entails much noise.

Table 5 presents results obtained from calculating the precision, recall, and F-value for the deflation determination accuracy using the method reported by Ebina et al. The number of moving averages to be monitored is applied to equation (4) for $n = 5$, $n = 10$, and $n = 15$. The results representing deflation are compared. Results showed that the judgment accuracy was best when $n = 10$.

Figure 5   Judging congestion from tweet information (see online version for colours)
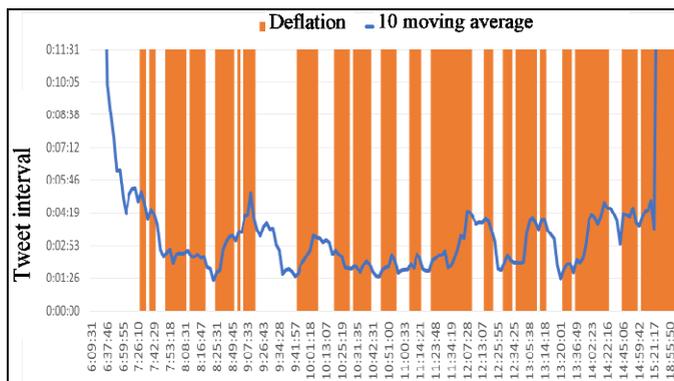

Table 5   Quantitative evaluation according to Ebina et al. (see online version for colours)

	Precision	*Recall*	*F-value*
5 moving average	25.77%	47.62%	33.44%
10 moving average	33.33%	55.24%	41.57%
15 number analysis	31.61%	52.38%	39.43%

4.4   Experimental results of our proposed method

The results obtained using the proposed method are shown in Figure 6. The broken line in the graph represents the change in the tweet posting time interval. The grey part shows the deflation judgment result obtained using the proposed method. The time at which the value of s in equation (1) is negative was defined as a deflation judgment condition.

Figure 6   Results of deflation determination using the proposed method (see online version for colours)
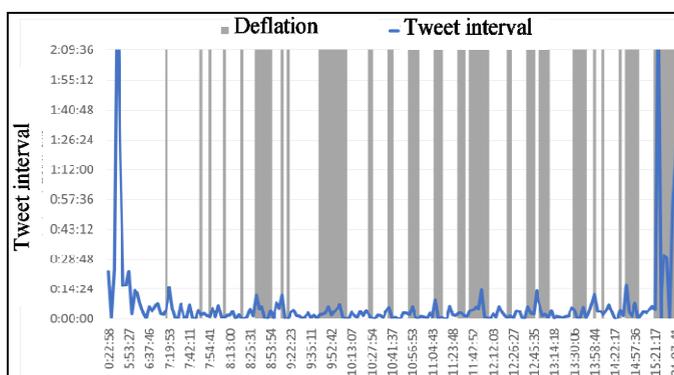
From the appearance time, we chose the correct time period during which people were thought to have moved, as shown in Table 3. Some time periods were estimated incorrectly as deflation, but there were also many time periods that were estimated fairly accurately: around 10:00 and around 12:00. The end of the event after 15:00 can be estimated correctly.

Table 6 presents results obtained from calculating the precision, recall, and F-value for the deflation determination accuracy using the proposed method. Comparing the number of tweets used for judgments as 6, 10 and 16, 10 was superior.

Table 6   Quantitative evaluation (see online version for colours)

	Precision	*Recall*	*F-value*
Using 6 (3 + 3) tweets	26.15%	26.98%	26.56%
Using 10 (5 + 5) tweets	52.63%	46.30%	49.89%
Using 16 (8 + 8) tweets	48.51%	45.37%	46.89%

4.5 *Summary of comparative experiments*

To evaluate the deflation results of this time quantitatively, a comparison was made using the precision, recall, and F-value. For correct answer data, the correct answer is found to be deflation during the time period shown in Table 3. Data to be compared were the experimentally obtained results using the technique of Ebina et al. (2012) and the experimentally obtained results using the technique of Endo et al. Table 7 presents comparison results.

The method of Endo et al. (2016) using tweet frequency had a good precision rate, but a poor recall rate. This method is useful for daily estimates, such as when to see cherry blossoms, but real-time estimates in minutes do not provide high accuracy.

The method reported by Ebina et al. using the tweet posting time interval yields the best recall because of its real-time nature. However, even if the reaction is good and no deflation is shown, a situation can be determined erroneously as deflation. Its precision is not always high.

Table 7   Quantitative evaluation (see online version for colours)

	Precision	*Recall*	*F-value*
Method of Endo et al. frequency / moving average	64.81%	25.18%	36.27%
Method of Ebina et al. tweet interval / moving average	33.33%	55.24%	41.57%
Proposed method tweet interval / number analysis	52.63%	46.30%	49.89%

Comparison of the relevance rates showed that the method proposed in this study provided the most balanced good results. Certainly, the precision rate is inferior to that of Endo et al. and the recall rate is inferior to that of Ebina et al. However, neither the precision nor the recall is extremely bad. The proposed method has the best accuracy when compared by F-value.

5 **Location estimation**

We will challenge to increase the dataset used to improve the accuracy of deflation judgment. In the previous chapter we estimated people's movements using tweets with geotag 'coordinates' that can obtain accurate location information. However, only 216 accounts posted the tweets used in the experiment. The number of visitors conducting a general visit was 210,000. 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 location. Therefore, we will perform verification by increasing the number of analysis targets using tweets with unclear location. We will use machine learning to extract the tweets posted at the target event venue from the group of tweets containing 'place' that represents Chiyoda ward. Estimate the posting location. The number of tweets to be analysed is therefore increased. Then the accuracy of deflation determination is evaluated quantitatively. First, we discuss extraction of tweets posted during the general visit from tweets that are clearly posted in Chiyoda ward.

The geotag 'place' has a rectangular range information represented by four latitudes and longitudes. This rectangle represents the location information at the city level. Compared to the number of tweets that have the geotag 'coordinates' that is accurate location information, the number of tweets that contain only the geotag 'place' is larger. It depends on the location, but there is a difference of about four times in Japan as a whole. The tweets that are useful as analysis targets are estimated from the group of tweets to which only the geotag 'place' was added. Those tweets have ambiguous location information, but they are added as analysis targets. Similarly to the case of analysis using only tweets with geotags 'coordinates', recall, precision and F-value are calculated. Then their accuracy is evaluated quantitatively.

5.1 'Place' data

As a tweet with a geotag for which location information is ambiguous, a tweet for which the geotag metadata is 'place_type:city' is used. The data to be classified by machine learning are tweets with no geotag 'coordinates' added on 4 May 2019. Only 'place' data representing Chiyoda ward, Tokyo are used. The Imperial Palace is located there. There were 3,132 tweets.

The attached 'place' data were confirmed for the 198 tweets with 'coordinates' used in the previous 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 7 portrays these four points. The range including this place is the range surrounding the entire Chiyoda ward.

Tweets that have only the geotag 'place' and no the 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 tweeted location cannot be found. Therefore, we analyse the tweet contents using natural language processing and consider a method to estimate the user's location more accurately based on the tweet contents. To infer the location, 3,132 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 to the Imperial Palace. When classifying tweets with only geotag 'place' added by machine learning, teacher data are extracted from text data of tweets with geotag 'coordinates' added. In the classification, the target data for estimating the location were text data of the tweet to which only the geotag 'place' representing Chiyoda ward was added. We limited to tweets posted on 4 May 2019, when the target event was held.

Figure 7   'Place' of Chiyoda ward (see online version for colours)
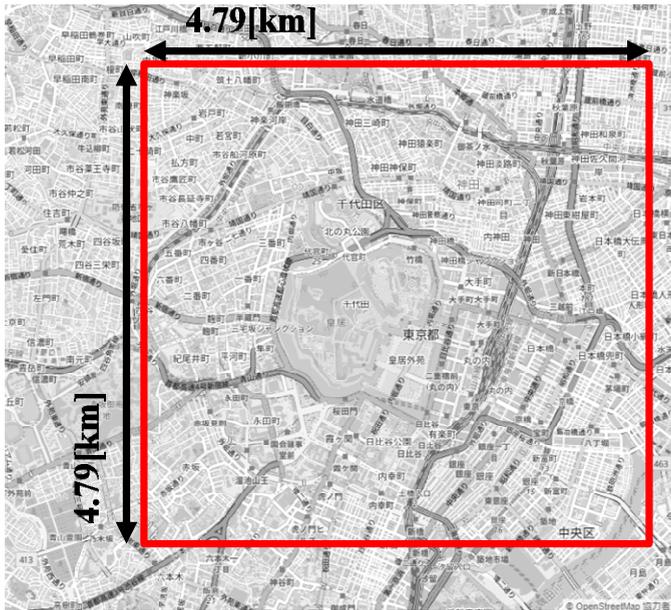
5.2   *Vectorisation and dimension reduction*

For using 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 analysing the tweet set morphologically and dividing it into words. This time, we extracted only nouns. Then, considering that the absolute number of tweets with the geotag 'coordinates' is small, we particularly examine nouns for each tweet. MeCab is used to extract the morphemes. Furthermore, by normalisation, 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 analyser, 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 vectorised. Bag of words (BoW) is used for vectorisation. With BoW, the number of times a word appears is counted for each tweet. A matrix is generated from the counted number of words.

The machine learning result is affected by vectorising the words extracted using morphological analysis. There is a great amount of noise if you simply vectorise without considering the importance and meaning of words. Two pre-processings, TF-IDF and LSI, were done to use the feature vector extracted from each tweet content as optimum data for use in machine learning.

The vectorised 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. Also, 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, which is a learning model capable of 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 generalisation performance. The following ranges for extracting correct and incorrect data are both included in the 'place' range (Figure 7) for Chiyoda ward.

Among the teaching data, the 198 tweets used in the experiment in the previous 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. Table 8 presents an example of a tweet. Latitude and longitude are included in the area depicted in Figure 3. Most tweet texts include the words '一般参賀 (general public)' and '皇居 (Imperial Palace)' related to the event. However, some tweets include no such characteristic words.

**Table 8** Examples of tweets posted in the target area

create_at	Longitude	Latitude	Text
11:31:15	139.755	35.6787	#一般参賀 in 千代田区, 東京都
13:20:40	139.755	35.6836	天気良過ぎ… in 千代田区, 東京都
13:30:06	139.755	35.6787	いよいよ (@ 皇居正門 in 千代田区, 東京都)
13:36:49	139.758	35.6787	I'm at 皇居前広場 in 千代田区, 東京都

Incorrect answer data was extracted from tweets in areas where the event did not occur, not including the Imperial Palace. These tweets have a 'coordinated' geotag. The range of the incorrect answer data is the same area as that of the correct answer data. This is the location of the target event, 200 meters east of the Imperial Palace. The range of latitude and longitude from the southwest is given below. Figure 8 portrays the 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]

Figure 8   Area of incorrect data (see online version for colours)
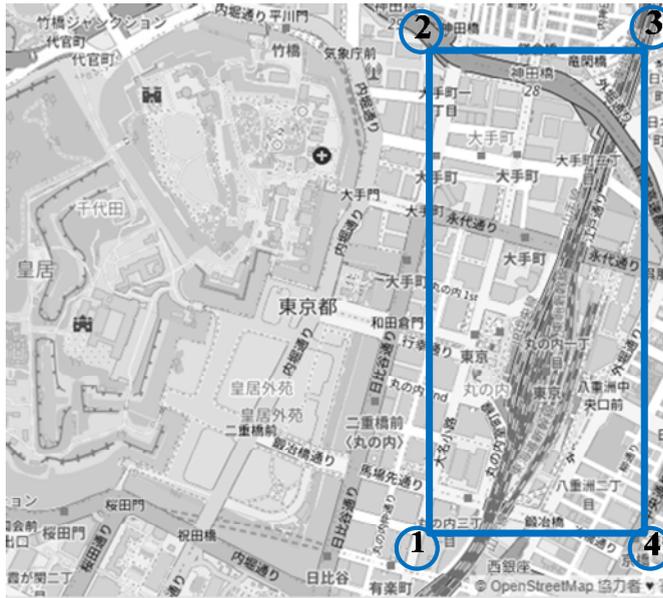
By extracting the tweets to which 'coordinates' were added under the conditions depicted in Figure 8, 866 tweets were eventually extracted. Then the text set that combines correct and incorrect data above is used as a learning model. Table 9 presents an example of a tweet. Latitude and longitude are included in the area shown in Figure 8. This area is not the target event occurrence area. However, there were 17 tweets containing the word '一般参賀 (general public)' and 13 tweets containing the word '皇居 (Imperial Palace)'. Most tweet texts include 'Tokyo station', which exists in this area, but some tweets do not include characteristic words.

Table 9   Examples of tweets posted outside the target area

create_at	Longitude	Latitude	Text
05:41:12	139.767	35.6809	旅に出ます探さないでください (@ 東京駅 in 千代田区, 東京都)
09:44:50	139.768	35.6799	皇居混みすぎ in 千代田区, 東京都
10:08:38	139.770	35.6800	#一般参賀 場所: Tokyo, Japan
12:38:22	139.770	35.6800	平和かも。 場所: Tokyo, Japan

5.4   SVM

We used a classification method by support vector machine (SVM) capable of binary classification location of a person who was at the scene of a specific event or trouble from the post to which only the geotag 'place' was added. Actually, SVM is adopted as a method to estimate the user attributes and location from tweet contents in many previous studies. 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 maximises 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 is 198 tweets with geotag 'coordinates' extracted in the correct answer range (Figure 3), which is regarded as having caused people to stagnate and flow because of the occurrence of events. The other is 866 tweets with geotag 'coordinates' extracted in the range (Figure 7) where the target event (visit of the General Public to the Imperial Palace after the accession to the throne) has not occurred. The 1,064 tweet data, which include these two types of data, 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 total. 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 (6).

$$
\begin{aligned}
&\textit{Classification accuracy rate} \\
&= (\textit{number of successful classifications})/(\textit{number of evaluation cases})
\end{aligned}
\tag{6}
$$

As a result, it was 80.64%.

5.5 *Estimated result*

The classification target is 3,132 tweets with only the geotag 'place' that represents Chiyoda ward. There are 184 tweets containing the word '皇居 (Imperial Palace)' that characterises the target event, and 231 tweets containing '一般参賀 (general public)'. However, tweets containing those feature words also existed in the tweets posted in the area presented in Figure 8. It is difficult to estimate whether a tweet was posted near the target area based on the presence or absence of feature words. Therefore, we use BoW to represent the tweet text as a 100-dimensional vector and perform location estimation by SVM. 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, 1,750 tweets were output as correct answers.

These 1,750 tweets are combined with data of 198 correct tweets with the geotag 'coordinates'. Using 1,948 tweets, we conduct analysis using our proposed method. By calculating the precision, recall, and F-value, the accuracy of deflation judgment can be evaluated quantitatively when analysis is performed by adding tweets estimated as having been posted in the range presented in Figure 3 by machine learning.

As in Table 6, the three types of parameters are shown. In Table 6, we evaluated by only tweets have the geotag 'coordinates'. But in Table 10, as a result of location estimation, the number of tweets used for evaluation is 1,948, which is larger than in Table 6. The number of tweets has increased, but the precision and recall have not improved.

Table 10   Quantitative evaluation using 1,948 tweets

	Precision	*Recall*	*F-value*
Using 6 (3 + 3) tweets	39.26%	32.73%	35.70%
Using 10 (5 + 5) tweets	46.35%	31.50%	37.51%
Using 16 (8 + 8) tweets	54.90%	37.24%	44.38%

We were able to increase the number of tweets used to judge deflation by machine learning from 198 to 1,948, nearly ten times. The tweet posting location was estimated using SVM, but its classification accuracy rate in cross validation is about 80%, and it contains about 20% noise. Even if the number of tweets used to judge deflation is increased by about 10 times, noise is included. Therefore, it was not possible to improve the accuracy of deflation judgment. Since our proposed method is delicate, even if the number of tweets to be analysed increases, the accuracy of deflation estimation will not improve if noise is included.

6   Conclusions

We proposed new method of analysing geotagged tweets for real-time examination of a crowded situation. We aimed to detect not only the crowded burst phenomenon but also its convergence. Real-time performance was realised. We succeeded in correct detection of the deflation phenomenon. Furthermore, the precision and F-value of the proposed method are superior to other methods.

SVM location estimation increased the number of tweets posted at the target event location. The classification accuracy rate indicated by five-fold cross validation performed only with tweets with the geotag 'coordinate' was not bad. However, due to insufficient location estimation accuracy, our proposed method, which is sensitive to noise, could not improve the judgment accuracy.

Increased versatility must be achieved as a deflation estimation method. Since few tweets can pinpoint the exact location of the post, we need to be able to more accurately estimate deflation even with tweets that have only ambiguous location information. Also, the method must be adaptable to various events. In the deflation estimation method, a general average value was used as the threshold value. Some other value might be more suitable as the threshold than the average value. Comparisons and investigations must be done using greater amounts of data. This is left as another issue for future work.

Acknowledgements

References

Ebina, R., Nakamura, K. and Oyanagi, S. (2010) 'A proposal for a real-time burst detection method', *DBSJ Journal*, Vol. 9, No. 2, pp.1–6.

Ebina, R., Nakamura, K. and Oyanagi, S. (2012) 'A proposal for a real-time burst analysis method', *IPSJ TOD*, Vol. 5, No. 3, pp.86–96.

Endo, M., Hirota, M., Ohno, S. and Ishikawa, H. (2017) 'Best-time estimation method by region and tourist spot using information interpolation', *IWIN2016*, pp.209–216.

Endo, M., Shoji, Y., Hirota, M., Ohno, S. and Ishikawa, H. (2016) 'On best time estimation method for phenological observations using geotagged tweets', *IWIN2016*, pp.205–210.

Kleinberg, J. (2002) 'Bursty and hierarchical structure in streams', in *Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp.91–101.

Morikuni, T., Yoshida, M., Okabe, M. and Umemura, K. (2015) 'Geo-location estimation of tweets with stop words detection', *Journal of Information Processing, Transaction on Database*, Vol. 8, No. 4, pp.16–26.

Ozer, O., Heri, R. and Kjetil, N. (2018) 'Locality-adapted kernel densities for tweet localization', *SIGIR'18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp.1149–1152.

Sakaki, T. and Matsuo, Y. (2012) '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.

Shasha, D. and Zhu, Y. (2004) 'High performance discovery in time series', *Techniques and Case Studies (Mono-Graphs in Computer Science)*, Springer-Verlag, New York.

Zhang, X. and Shasha, D. (2006) 'Better burst detection', *Proc. 22nd International Conference on Data Engineering*, IEEE Computer Society, pp.146–149.

Zhu, Y. and Shasha, D. (2003) 'Efficient elastic burst detection in data streams', *Proc. 9th ACM SIGMOD International Conference on Knowledge Discovery and Data Mining*, ACM, pp.336–345.