2016 5th IIAI International Congress on Advanced Applied Informatics

# Location-based Temporal Burst Detection using Outlier Factors in Geo-tagged Tweets

Keiichi Tamura†, Tatsuhiro Sakai‡ and Hajime Kitakami†
Graduate School of Information Sciences, Hiroshima City University,
3-4-1, Ozuka-Higashi, Asa-Minami-Ku, Hiroshima 731-3194, Japan
E-mail:†{ktamura,kitakami}@hiroshima-cu.ac.jp,‡da65003@e.hiroshima-cu.ac.jp

***Abstract*—Burst detection is one of the most popular techniques for extracting remarkable keywords in online social documents posted through social media. With the growing interest in geosocial media these days, many researchers are focusing on extracting geolocal keywords related to local topics and events from such social documents because of the increasing number of geo-annotated documents (e.g., geo-tagged tweets on Twitter). In our previous work, we proposed a method for identifying local temporal burstiness to detect local hot keywords considering a user's location. Our method is based on moving-average convergence/divergence(MACD)-histogram-based temporal burst detection, and the experiments indicated that our method can sensitively identify the local temporal burstinesses of keywords on the basis of public awareness. However, daily fluctuations in the occurrence rates of keywords have an effect on the qualities of location-based burst detection. To tackle this issue, we propose a new method for identifying local temporal burstiness using quartile-based outlier factors in this paper. Utilization of the quartile-based outlier factors allows for location-based burst detection to eliminate daily fluctuation phenomena. To evaluate the proposed method, we conducted experiments using actual geo-tagged tweets posted on Twitter. The experiments revealed that the proposed method can identify local temporal burstiness more sensitively compared with our previous method.**

***Keywords*—*Burst detection, Geo-tagged tweet, Social media, Time series analysis, Outlier detection***

## I. Introduction

With the advent of the era of big data, the extraction of social topics and events that attract people's attention in social contents posted on social media sites has become more popular. People post information about things that they witness through contents such as messages, photographs and videos. Therefore, the contents on social media sites have a large amount of potential for the analysis of public awareness in real time. Twitter is one of the most well-researched social media sites. People post not only personal topics but also social topics and events related to geolocation. Bursty keywords appearing in tweets allow us to detect social topics and events because tweets posted on Twitter can be assumed to be a document stream.

Burstiness has been one of the most important criteria for extracting topics and events in document streams [1]. If a topic or an event attracts people's attention, the number of documents including keywords related to the topic or the event rapidly increases, and this phenomenon is called a burst. A significantly large number of studies has been conducted on burst detection and its applications. Detection of the burstinesses of keywords appearing in tweets helps to extract interesting social topics and events. This opened up a new era of media in the sense that people obtain information from social media.

Through the spread of smartphones equipped with a global positioning system, a large number of geo-tagged social contents is posted on social media sites [2]. For example, geo-tagged tweets posted on Twitter include not only text messages but also geolocation information such as the longitude and latitude. This provides a new challenge of extracting temporal bursty keywords closely related to local topics and events. For example, if the local topic "TA" attracts people in the region "RA," users in the region "RA" post geo-tagged tweets mentioning "TA." This phenomenon increases the degree of temporal burstiness of "TA" in "RA."

To tackle this issue, we have proposed a location-based temporal burst detection algorithm [3], [4]. In our previous work, we proposed a method for detecting local temporal burstiness by utilizing moving-average convergence/divergence(MACD)-histogram-based temporal burst detection algorithm [5] that can identify the local temporal burstiness of keywords considering a user's location. Our previous method can detect local bursty periods of keywords and contributes to a region-dependent extraction of bursty keywords; however, routine natural fluctuations in the occurrence rates of keywords have an effect on the qualities of burst detection.

In this study, we propose a novel method for detecting location-based burstiness using outlier factors. Our new method utilizes the quartile-based outlier factor to decrease the influence of routine occurrence rates of keywords. In particular, our method detects how large the total amount of the distance-based importance rates of keywords raises from usual state. The remainder of the paper is organized as follows. In Section II, the related work is reviewed. In Section III, MACD-histogram-based local temporal burst detection is briefly described. In Section IV, we propose a new method for identifying local temporal burstiness. In Section V, experimental results are reported, and Section VI concludes the paper.

## II. Related Work

Nowadays, geo-annotated social contents are becoming one of the most influential information sources in the world. For extracting topics and events from social contents, a large number of researchers in different application domains has attempted to accelerate and enhance the use of geo-annotated social contents. For example, Sakaki et al. [6] proposed a

978-1-4673-8985-3/16 $31.00 © 2016 IEEE
DOI 10.1109/IIAI-AAI.2016.155

novel method for detecting earthquake occurrences using geo-tagged tweets from Twitter. They considered each Twitter user as a social sensor and developed a method for estimating the centers of earthquakes. Arase et al. [7] remarked that frequent trip patterns in geo-tagged photographs posted for sharing on social media sites could be beneficial for tourism industries in that such information can assist tourism companies when recommending places of interest. Since Kleinberg proposed Kleinberg's temporal burst detection algorithm, which has had a very significant influence on many studies, burst detection has been one of the most thoroughly studied topics for detecting topics and events that attract people on the Internet from online documents. Kleinberg's temporal burst detection algorithm [1] is based on queuing theory for detecting bursty network traffic and can be used to analyze document streams from various sources such as e-mails [1], blogs [8], online publications [9], bulletin boards, and social tags [10].

Researchers have started to study the temporal bursts of local topics and events because of the increasing number of georeferenced documents. Michael et al. [11] proposed a new methodology to identify spatial bursts, whereby the temporal interval of interest is given preliminarily. Their work focused on identifying geographical regions where the observed frequency of terms was higher than usual. In [12], for each region, a local document stream was used and a new method for finding temporal burstiness patterns among several regions was devised. A search engine that considers the spatiotemporal burstiness of the terms in the process of document retrieval was demonstrated. Zimmermann et al. [13] presented a clustering method that detects, tracks, and updates large and small bursts of news in a two-level topic hierarchy. All of these studies anticipated the development of techniques for spatiotemporal burstiness. They did not address the need to identify location-based temporal bursts.

## III. MACD-HISTOGRAM-BASED LOCAL TEMPORAL BURST DETECTION

This section reviews the MACD-histogram-based local temporal burst detection algorithm proposed in [14].

### A. Data Model

Suppose that a keyword $key$ is considered as a subject of supervision. In this study, the local temporal burstiness of $key$ is measured every hour, and a set of geo-tagged tweets is created hourly. A set of geo-tagged tweets is called a batch of geo-tagged tweets. Let a time series of batches be $SBGT = (BGT_1, BGT_2, \cdots, BGT_n)$. Let $BGT_i$ be the $i$-th batched geo-tagged tweets in $SBGT$: $BGT_i = \{gt_{i,1}, gt_{i,2}, \cdots, gt_{i,numgt(i)}\}$, where $numgt(i)$ denotes the number of geo-tagged tweets in the $i$-th batch. Fig. 1 shows an example. In this example, there are five batches, and each batch of geo-tagged tweets is created every hour.

The geo-tagged tweet $gt_{i,j}$ consists of three items, namely text data, the post time, and location information: $gt_{i,j} =< msg_{i,j}, pt_{i,j}, pl_{i,j} >$, where $msg_{i,j}$ denotes the text message, $pt_{i,j}$ represents the post time, and $pl_{i,j}$ refers to the location at which $gt_{i,j}$ was posted or is located. In this study, longitude and latitude coordinates are used for location information. For example in Fig. 1, there are five geo-tagged tweets in $BGT_3$ that are mapped onto the geographical coordinate space.

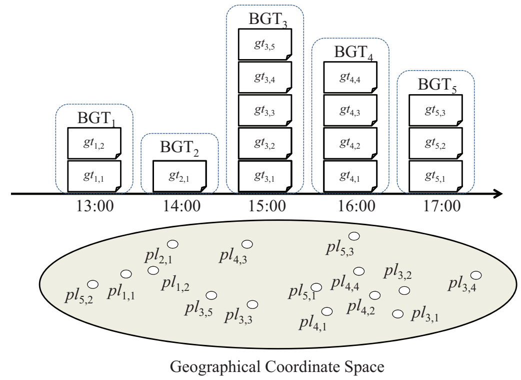



Fig. 1: Data model.

### B. Local Temporal Burstiness

Suppose that it rains heavily in a particular area "RA." In this case, the keyword "rain" becomes a temporal bursty keyword in the area "RA," and it is a hot topic near it. However, "rain" is not an important topic for users located far from "RA" if it does not rain in the area where the users are located. In this case, the keyword "rain" should be presented as a highly temporal bursty topic for users close to area "RA," whereas it should not be presented as a highly temporal bursty topic for users far away from this region. Location-based burst detection attempts to extract local temporal burstiness for detecting local hot keywords.

### C. MACD and Its Burst Detection

To consider the users' locations, the distance-based importance rates of geo-tagged tweets, which are determined by their distances from a user, are integrated into the MACD-histogram-based burst detection algorithm. In this study, forgetting theory [15] is used to calculate the distance-based importance rates. Geo-tagged tweets gradually lose their importance as the distance from the user increases as

$$infr_{t,j} = \beta^{distance(pl_{t,j}, ul)}, \tag{1}$$

where the function $distance$ returns the distance between a user's location $ul$ and $pl_{t,j}$, and $\beta$ ($0.0 < \beta < 1.0$) represents the forgetting factor. Fig. 2 shows an example of the distance-based importance rates. The distance-based importance rate of $gt_{4,1}$ is 0.9 for $User$ 1, whereas that of $gt_{1,1}$ is 0.1 because $gt_{1,1}$ is far away from $User$ 1. On the contrary, the distance-based importance rate of $gt_{4,1}$ is 0.1 for $User$ 1, whereas that of $gt_{1,1}$ is 0.9.

The following is the definition of the total amount of the distance-based importance rates in $BGT_t$:

$$ibgt_t = \sum_{j=1}^{|BGT_t|} infr_{t,j}. \tag{2}$$

Let a time series of the total amount of the distance-based importance rates in $BGT_t$ be $sibgt = (ibgt_t \mid t = 1, 2, ...)$.

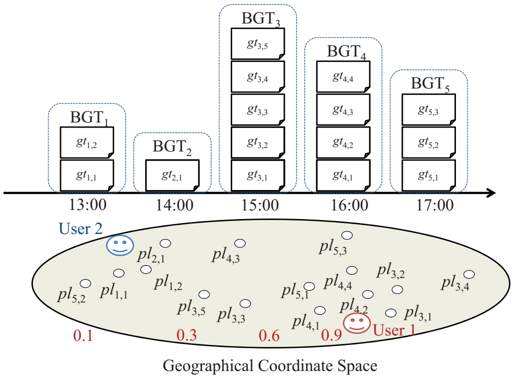



Fig. 2: Distance-based importance rates.

This time series is used for the MACD-histogram-based temporal burst detection algorithm. He et al. [5] regarded the time series of the number of documents including a keyword as a two-dimensional position. Location-based temporal burst detection is based on this time series of the total amount of the distance-based importance rates.

The definition of the exponential moving average (EMA) for the $t$-th items of $sibgt$ is

$$\begin{aligned} EMA(n)[sibgt]_t &= \alpha \times ibgt_t + (1-\alpha)EMA(n)[sibgt]_{t-1} \\ &= \sum_{k=0}^{n} (\alpha(1-\alpha)^k ibgt_{t-k}), \end{aligned} \quad (3)$$

where $n$ is the window size, and $\alpha = 2/(n-1)$.

The difference between $EMA(n_1)[sibgt]_t$ and $EMA(n_2)[sibgt]_t$ is called the MACD, where $n_1 \neq n_2$:

$$\begin{aligned} &MACD(n_1, n_2)[sibgt]_t = \\ &EMA(n_1)[sibgt]_t - EMA(n_2)[sibgt]_t, \;\; n_1 > n_2. \end{aligned} \quad (4)$$

The MACD is considered to be a derivative value of the EMA and is a velocity. This was introduced by Gerald Appel in the late 1970s [16] and is used in the technical analysis of stock prices. Therefore, $MACD(n_1, n_2)[sibgt]_t$ is a velocity of $sibgt$.

The EMA of the MACD, where the window size is $n_3$, is called the MACD signal. The difference between the signal and the MACD is called the MACD histogram. The MACD histogram is a derivative value of the MACD and is regarded as the acceleration of the time series.

$$\begin{aligned} &signal(n_1, n_2, n_3)[sibgt]_t = \\ &\quad EMA(n_3)[MACD(n_1, n_2)[sibgt]]_t, \end{aligned} \quad (5)$$

$$\begin{aligned} &histogram(n_1, n_2, n_3)[sibgt]_t = \\ &\quad MACD(n_1, n_2)[sibgt]_t - signal(n_1, n_2, n_3)[sibgt]_t, \end{aligned} \quad (6)$$

where $MACD(n_1, n_2)[sibgt]$ denotes the time series of the MACD values for $sibgt$.

If $histogram(n_1, n_2, n_3)[sibgt]_t > 0$, this means that the derivative of the EMA is positive and the period of the $t$-th item is a local temporal bursty period. Otherwise, the period of the $t$-th item is not a local temporal bursty period.

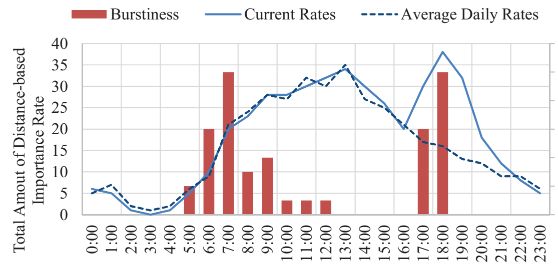



Fig. 3: Problem definition.

## IV. Proposed Method

In this section, the novel location-based temporal burst detection algorithm is proposed.

### A. Problem Definition

The frequencies of several keywords commonly appearing in geo-tagged tweets regularly increase and decrease. For example, the occurrence rate of the keyword *"good morning,"* rapidly increases in the morning and then rapidly decreases. If we do not consider the movement of the regular occurrence rate, this phenomenon is detected as a temporal bursty event. Therefore, we must not measure MACD histograms but also take into account the regular alternations of the occurrence rates of keywords appearing in geo-tagged tweets.

Fig. 3 shows a detailed example of our motivation. The labels "Current Rates" and "Average Daily Rates" are time series of the total amount of the distance-based importance rates and the average daily total amount of the distance-based importance rates respectively. Bars show the degrees of Burstiness based on "Current Rates." There are two bursty periods; however, the first temporal bursty period from 5:00 to 12:00 appears regularly. On the other hand, the second temporal bursty period from 17:00 to 18:00 can be a sudden topic because the total amount of the distance-based importance rates rapidly increases compared with the usual rates. In this example, we only have to extract the second temporal bursty period.

### B. Quartile-based Outlier Factors

To address this issue, we integrate the quartile-based outlier scheme into location-based temporal burst detection. The quartile-based outlier is based on the disciplines of statistics and statistical analysis. Descending data are divided into four parts, and the quartiles are three break points. The first quartile denoted $Q_1$ is the median between the smallest value of the dataset and the median of the dataset. The second quartile denoted $Q_2$ is the median of the dataset and the third quartile denoted $Q_3$ is the median between the highest value of the dataset and the median of data. The difference between $Q_3$ and $Q_1$ is called the interquartile range (IQR); thus, $IQR = Q_3 - Q_1$.

In quartile-based outlier detection, there are two types of fences, the lower fence (LF) and upper fence (UP):

$$LF = Q_1 - 1.5 \times IQR, \; UP = Q_3 + 1.5 \times IQR. \quad (7)$$

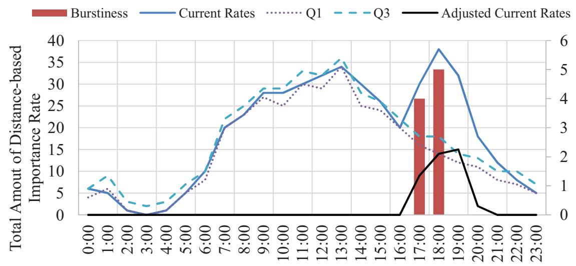



Fig. 4: An example of adjusting rates.

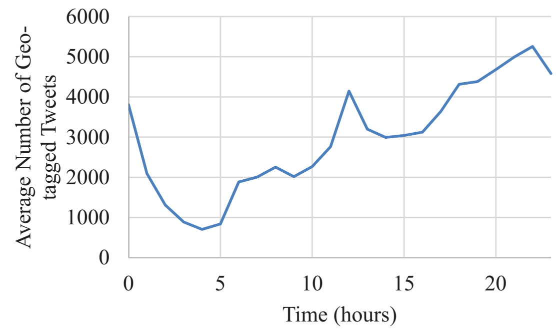



Fig. 5: Average number of geo-tagged tweets of "Training Set."

If a datum in the dataset is located outside these fences, this value is an outlier.

### C. Algorithm

To utilize the scheme for quartile-based outlier factors for detecting local temporal burstiness, the total amount of the distance-based importance rates every hour is measured for $m$ days using a training set and denoted by the matrix $IM$. The element $IM_{i,j}$ is the total amount of the distance-based importance rates at $j$ on the $i$-th day of the test set, where $i = \{1, 2, \cdots, m\}$, and $j = \{0, 1, \cdots, 23\}$.

Let the first, second, and third quartiles of the total amount of the distance-based importance rates every hour be $Q_{1,j}$, $Q_{2,j}$, and $Q_{3,j}$, respectively. The IQR is the difference between $Q_{3,j}$ and $Q_{1,j}$; thus, $IQR_j = Q_{3,j} - Q_{1,j}$, and it is a factor of the degree of the distribution in the $j$-th column of $IM$. For example, suppose that the $k$-th column of $IM$ is $IM_k^T = (1.7, 2.8, 1.0, 2.9, 9.0, 1.5, 1.4)$. The values of $Q_{1,k}$, $Q_{2,k}$, and $Q_{3,k}$ are 1.4, 1.7, and 2.9 respectively. The interquartile range is $IQR_k = 1.5$.

The following is the function for detecting an outlier:

$$IsOutlier(x, q_1, q_3) = \begin{cases} true, \ if \ x \geq 1.5 \times (q_3 - q_1) + q_3, \\ false, \ otherwise. \end{cases} \tag{8}$$

The total amount of the distance-based importance rates $ibgt_t$ of $BGT_t$ is adjusted as follows:

$$aibgt_t = \begin{cases} ibgt_t - 1.5 \times (Q_{3,\phi(t)} - Q_{1,\phi(t)}) - Q_{3,\phi(t)}, \\ \quad if \ IsOutlier(ibgt_t, Q_{1,\phi(t)}, Q_{3,\phi(t)}) = true \\ 0, \quad otherwise, \end{cases} \tag{9}$$

where $\phi(t)$ returns the hour that $BGT_t$ is posted. Let a time series of the adjusted distance-based importance rates be $saibgt$. If $histogram(n_1, n_2, n_3)[saibgt]_t > 0$, the period of the $t$-th item is a temporal bursty period. Otherwise, the period of the $t$-th item is not a temporal bursty period.

Fig. 4 shows an example of a time series of the adjusted distance-based importance rates. In this graph, the lines "Q1" and "Q3" are the first and third quartiles of the daily distance-based importance rates of $key$ for the user. The line denoted as "Adjusted Current Rates" is a plot of the adjusted distance-based importance rates. The adjustment of the rates using the UP helps to detect actual temporal burst phenomena.

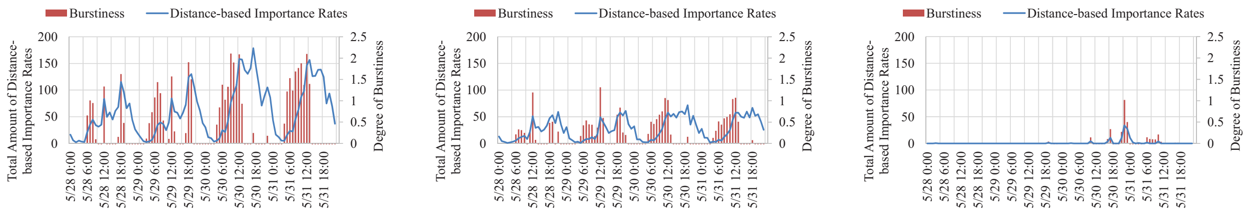
## V. Experiments

To evaluate the proposed location-based burst detection algorithm, we collected geo-tagged tweets from Twitter using its streaming application program interfaces (APIs). We used two datasets: one was a set of geo-tagged tweets (denoted by "Training Set") from May 7 to May 28, 2015, and the other was a set of geo-tagged tweets (denoted by "Test Set") from May 29 to May 31, 2015. "Training Set" was used for measuring the daily total amount of the distance-based importance rates. Fig. 5 shows the average number of geo-tagged tweets of "Training Set." For each keyword, we calculated the quartiles of the keyword using premeasured rates. "Test Set" was used to detect location-based burstiness. Two cities in Japan, namely, Tokyo and Osaka, were set as the users' locations. In the experiments, $n_1$, $n_2$, $n_3$, and $\beta$ were set to 3, 5, 4, and 0.9, respectively.

For each experiment, keywords were ranked by the total amount of burstiness, and top-3 bursty keywords were extracted for evaluation. The top-3 bursty keywords of our previous method are "Tokyo," "Shop," and "Sta" when the user's location is set to Tokyo. The top-3 bursty keywords of the proposed method are "Earthquake," "Tokyo," and "Sta" when the user's location is set to Tokyo. Fig. 6a, 6b, and 6c show the results for "Tokyo," "Shop," and "Sta." Fig. 7a, 7b, and 7c show the results for "Earthquake," "Tokyo," and "Sta." In each graph, "Burstiness" is a time series of the degree of burstiness and "Distance-based Importance Rates" is a time series of the total amount of the distance-based importance rates for the user.

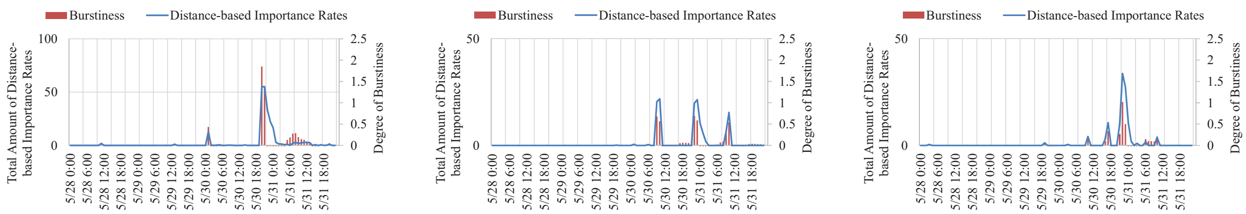
The keyword "Tokyo" is the top frequent term in Tokyo. Therefore, the occurrence rate is small at midnight and increases during the daytime. This daily routine phenomenon has a considerable effect on detecting bursts. Fig. 10 shows the distance-based importance rates of "Tokyo" and the quartiles of "Training Set." The burst phenomena of this keyword repeat daily, and we find that the increase in the occurrence rate of the keyword "Tokyo" is a daily phenomenon and not a bust. The keyword "Shop" is the same as the keyword "Tokyo."

At the same time, the first bursty keyword is "Earthquake," when we use the proposed method. A big earthquake hit the Tokyo region on May 30, 2015. This earthquake was a deep-focus earthquake, and a very large number of people posted



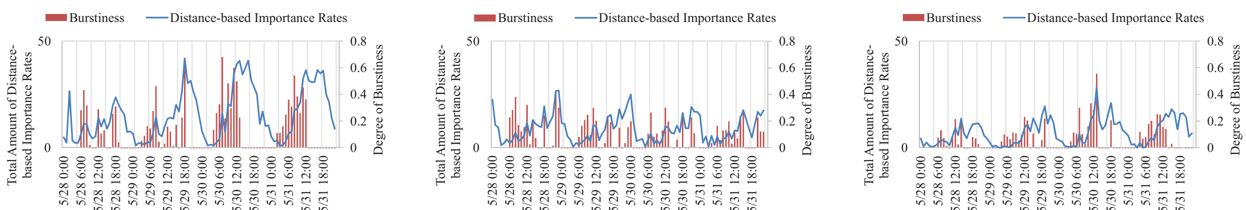

(a) Tokyo (b) Shop (c) Sta

Fig. 6: Results for the top-3 bursty keywords using the previous method in Tokyo.



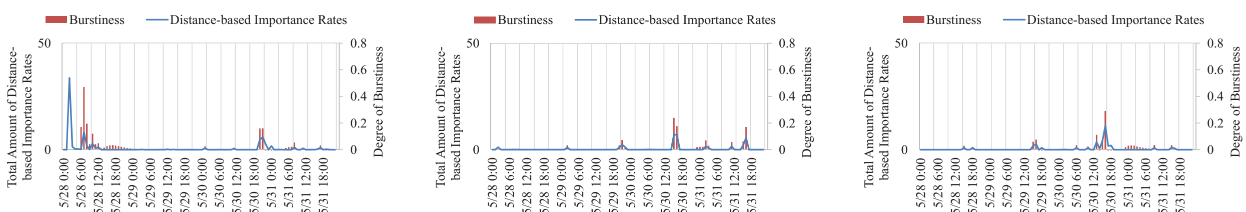

(a) Earthquake (b) Tokyo (c) Sta

Fig. 7: Results for the top-3 bursty keywords using the proposed method in Tokyo.




(a) Osaka (b) Smile (c) Shop

Fig. 8: Results for the top-3 bursty keywords using the previous method in Osaka.




(a) Earthquake (b) Osaka (c) Osaka Castle Park

Fig. 9: Results for the top-3 bursty keywords using the proposed method in Osaka.

messages related this earthquake. This sudden topic was one of the most attractive topics between May 28 and May 31. The proposed method can extract the keyword "Earthquake" as the top topic in Tokyo. In the proposed method, the keyword "Tokyo" is the second bursty keyword because messages related to the earthquake included the keyword "Tokyo." The keyword "Tokyo" is also bursty in the morning. A Japanese idol group held an event in an international exhibition center. Many people reported arriving the nearest station of the center with its address.

The third bursty keyword obtained by the previous and the proposed methods is "Sta" (Fig. 6c and Fig. 7c). Both cases reach a peak on May 30, 2015 at night. After the earthquake, almost all train lines were suspended. Many people reported train situations; therefore, a large number of messages related to stations was posted. This phenomenon appears as a burst during this period.

The keyword "Osaka" is the first bursty keyword in Osaka when we use the previous method (Fig. 8a). Moreover, the keywords "Smile" and "Shop" are the second and third bursty keywords respectively (Fig. 8b and Fig. 8c). These keywords are frequent daily keywords. Fig. 11 shows the distance-based importance rates of "Osaka" and the quartiles of "Training Set." For both Tokyo and Osaka, the keyword "Shop" is

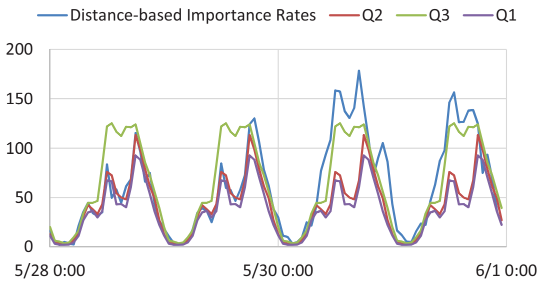



Fig. 10: Quartiles of the keyword "Tokyo."

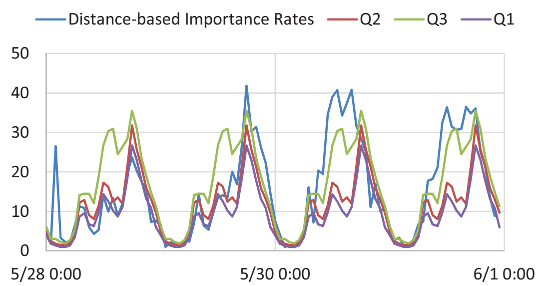



Fig. 11: Quartiles of the keyword "Osaka."

actively posted because many people write messages about the shops that they visit.

On the other hand, the first bursty keyword in Osaka, when the proposed method is used, is "Earthquake" (Fig. 9a). There are two peaks on May 28 and on May 30. The distance-based importance rates at midnight on May 28 are high, but there are no bursty. The MACD-histogram-based burst detection needs several data points. This period is the initial period; therefore, the proposed method cannot extract this period as a bursty period. The second bursty period occurs on May 30. There are no bursty periods for the keywords "Smile" and "Shop." The second and third bursty keywords are "Osaka" and "Osaka castle park" ( Fig. 9b and Fig. 9c). A popular band held a concert at the Osaka Castle Park, therefore, these two keywords are posted on May 30.

## VI. Conclusion

In this paper, we proposed a new method for identifying local temporal burstiness using quartile-based outlier factors. Utilizing the quartile-based outlier factors allows the location-based burst detection algorithm to eliminate daily fluctuation phenomena. To evaluate the proposed method, we conducted experiments using actual geo-tagged tweets posted on Twitter. The experiments revealed that the proposed method can identify local temporal burstiness more sensitively compared with our previous method. In our future work, we intend to develop an adaptive method considering the difference in posting rates.


## Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 26330139 and Hiroshima City University Grant for Special Academic Research.


## References


[1] J. Kleinberg, "Bursty and Hierarchical Structure in Streams," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '02, 2002, pp. 91–101.

[2] M. Naaman, "Geographic information from georeferenced social media data," *SIGSPATIAL Special*, vol. 3, no. 2, pp. 54–61, Jul. 2011.

[3] K. Tamura and H. Kitakami, "Detecting location-based enumerating bursts in georeferenced micro-posts," in *Proceedings of the 2013 IIAI International Conference on Advanced Applied Informatics (IIAIAAI)*, Aug 2013, pp. 389–394.

[4] S. Kotozaki, K. Tamura, and H. Kitakami, "Identifying burstiness of local topic using location-based burst detection with a classifier technique," in *Proceedings of the 2014 IEEE 7th International Workshop on Computational Intelligence and Applications (IWCIA)*, Nov 2014, pp. 225–230.

[5] D. He and D. S. Parker, "Topic dynamics: An alternative model of bursts in streams of topics," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '10, 2010, pp. 443–452.

[6] T. Sakaki, M. Okazaki, and Y. Matsuo, "Tweet analysis for real-time event detection and earthquake reporting system development," *IEEE Trans. on Knowl. and Data Eng.*, vol. 25, no. 4, pp. 919–931, Apr. 2013.

[7] Y. Arase, X. Xie, T. Hara, and S. Nishio, "Mining people's trips from large scale geo-tagged photos," in *Proceedings of the International Conference on Multimedia*, ser. MM '10, 2010, pp. 133–142.

[8] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, "On the bursty evolution of blogspace," in *Proceedings of the 12th International Conference on World Wide Web*, 2003, pp. 568–576.

[9] K. K. Mane and K. Börner, "Mapping topics and topic bursts in pnas," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101 Suppl 1, pp. 5287–5290, 2004.

[10] J. Yao, B. Cui, Y. Huang, and X. Jin, "Temporal and social context based burst detection from folksonomies," in *Proceedings of AAAI 2010*, 2010.

[11] M. Mathioudakis, N. Bansal, and N. Koudas, "Identifying, attributing and describing spatial bursts," *Proc. VLDB Endow.*, vol. 3, no. 1-2, pp. 1091–1102, Sep. 2010.

[12] T. Lappas, M. R. Vieira, D. Gunopulos, and V. J. Tsotras, "On the Spatiotemporal Burstiness of Terms," *Proc. VLDB Endow.*, vol. 5, no. 9, pp. 836–847, May 2012.

[13] M. Zimmermann, I. Ntoutsi, Z. F. Siddiqui, M. Spiliopoulou, and H.-P. Kriegel, "Discovering Global and Local Bursts in a Stream of News," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, ser. SAC '12, 2012, pp. 807–812.

[14] K. Tamura, T. Matsui, H. Kitakami, and T. Sakai, "Identifying local temporal burstiness using macd histogram," in *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, Oct 2015, pp. 2666–2671.

[15] Y. Ishikawa, Y. Chen, and H. Kitagawa, "An on-line document clustering method based on forgetting factors," in *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, 2001, pp. 325–339.

[16] G. Appel, *Technical Analysis: Power Tools for Active Investors*, 1st ed. FT Press, 2005.