International Journal of Electronic Commerce Studies
Vol.6, No.2, pp.269-288, 2015
doi: 10.7903/ijecs.1347

IDENTIFYING LOCAL BURSTINESS IN A SEQUENCE OF BATCHED GEOREFERENCED DOCUMENTS

Shota Kotozaki
Hiroshima City University
3-4-1, Ozuka-Higashi, Asa-Minami-ku, Hiroshima, 731-3194 Japan
my67010@e.hiroshima-cu.ac.jp

Keiichi Tamura
Hiroshima City University
3-4-1, Ozuka-Higashi, Asa-Minami-ku, Hiroshima, 731-3194 Japan
ktamura@hiroshima-cu.ac.jp

Hajime Kitakami
Hiroshima City University
3-4-1, Ozuka-Higashi, Asa-Minami-ku, Hiroshima, 731-3194 Japan
kitakami@hiroshima-cu.ac.jp

ABSTRACT

One of the most interesting emerging topics in social media is the increase in the number of georeferenced documents. These documents include not only text data, but also posted time and location. People have been transmitting information regarding items and events they have witnessed in their daily lives and collecting information on objects of interest through georeferenced documents. Therefore, many researchers are directing their attention to extracting local topics and events from georeferenced documents. In this paper, we propose a novel location-based burst detection algorithm for identifying the burstiness of a keyword related to local topics and events in a sequence of batched georeferenced documents, composed of ordered georeferenced document sets. Burstiness is one of the simplest yet most robust criteria for extracting hot topics and events from a sequence of batched documents. Identifying the burstiness of a keyword related to local topics and events captures not only the peak periods of the trending topics and events, but also the localities at which they are occurring. To evaluate the proposed location-based burst detection algorithm, we used an actual sequence of batched georeferenced documents that were composed by crawling tweets posted on the Twitter site. The

experimental results confirm that the proposed location-based burst detection algorithm can identify location-based bursts successfully.

Keywords: Burst Detection, Social Media, Georeferenced Document, Location-based Awareness

1. INTRODUCTION

In recent years, with the increasing attention directed towards social media and the wider use of smart phones, we now have access to an enormous number of short text messages posted on social media sites (e.g., Twitter and Facebook). The contents of these short text messages usually include not only personal daily information, but also social data on important topics and events. This trend for people to actively collect and transmit information to the world through these short text messages has accelerated in recent years. Furthermore, this emerging trend has formed a new type of media that can significantly influence people in the world. Therefore, extracting and tracking topics and events in from short text messages on social media sites represents an attractive research topic in many different applications domains.

On social media sites, we have witnessed the emergence of a new type of information that assumes the form of geo-annotated text messages. These are referred to as georeferenced documents[1, 2] in this paper. These georeferenced documents include posted time and location information as well as text data. Increasing interest in geo-location applications has led to an extensive number of georeferenced documents being posted by people on social media sites. Because social topics and events are often involved in local topics and events, these georeferenced documents are becoming a leading source of location information. This trend presents us with new research challenges such as how to identify when and where attractive local topics and events occur.

Burstiness is one of the simplest and most effective criterion that can be used to extract trending topics and events from georeferenced documents. The number of georeferenced documents related to a local topic or event increases sharply as the local topic or event attracts the interest of more people and, conversely, decreases sharply when interest in the local topic or event diminishes. We can detect trending topics and events by extracting bursty periods of keywords referencing the local topics or events. The most widely known burst detection algorithm for a sequence of batched documents is Kleinberg's[3]. Kleinberg's burst detection algorithm is designed to identify certain discrete periods in which the number of

documents including a keyword increases sharply compared with the normal situation. Kleinberg's burst detection algorithm is the simplest yet most notable algorithm for detecting bursts in a sequence of document sets. However, it does not consider the localities of the bursts associated with topics and events.

Suppose that it snows heavily in a particular area "A." In this case, the keyword "snow" becomes a bursty keyword in the area "A" and is a hot topic near the area "A." However, the keyword "snow" is not an important topic for users located distant from "A," if it does not snow in the area where the users are located. This example indicates that we must detect the locality of burstiness in consideration of users at various locations. The keyword "snow" should be presented as a highly bursty topic for users close to area "A," whereas it should be presented as not highly bursty for users distant from that region.

In this paper, we propose a novel algorithm for identifying local burstiness of a keyword, including local topics, in a sequence of batched georeferenced documents. The main contributions of this study can be summarized as follows.

- In our previous work[4], we proposed a location-based burst detection algorithm that identifies bursty phenomena of local topics and events in a sequence of batched georeferenced documents. The bursty phenomenon in a local topic occurs when the number of relevant georeferenced documents that include a keyword related to the local topic increase rapidly compared with usual conditions. In the location-based burst detection algorithm, the number of relevant georeferenced documents is adjusted using influence rates to detect the local burstiness according to the users' locations. The influence rate of a georeferenced document is determined by the distance between the georeferenced document and user. However, this definition does not reflect the relationships between georeferenced documents. This sometimes causes noisy bursts that are not related to local topics and events. For example, if some georeferenced documents are located in a small region, these georeferenced documents are strongly connected to local topics and events. Otherwise, these georeferenced documents are not strongly connected to local topics and events. Our previous algorithm did not consider this situation. To overcome this issue, we integrate density-based criteria in extracting location-based bursts more precisely.
- The proposed new location-based burst detection algorithm utilizes kernel density estimation to integrate the relationships between

georeferenced documents in location-based burst detection. However, kernel density estimation does not consider temporal changes in density. To consider temporal changes in density, we have modified the definition of kernel density estimation. The new definition is based on the time forgetting theory. In the time forgetting theory, influence decreases as the time interval increases.

- To evaluate the proposed algorithm for identifying the local burstiness of keywords appearing in a sequence of georeferenced documents, an actual sequence of batched georeferenced documents constructed by crawling tweets posted on the Twitter site was used. The experimental results confirm that the new location-based burst detection algorithm can identify the local burstiness of a keyword, including weather-related topics in Japan, more precisely than the conventional algorithm.

The remainder of the paper is organized as follows. In Section 2, related work is reviewed. In Section 3, we briefly explain Kleinberg's burst detection algorithm. In Section 4, the data model and the problem definition are presented. In Section 5, local burstiness and a new algorithm for location-based burst detection are described. In Section 6, an evaluation of the work is presented. We conclude the paper in Section 7.

2. RELATED WORK

Geo-annotated user-generated data on social media sites is becoming one of the most influential sources of information. Many researchers have attempted to accelerate and enhance the use of geo-annotated user-generated data. For example, Sakai et al.[5] proposed a novel method for detecting earthquake occurrences using geo-annotated tweets on the Twitter site. They consider each Twitter user as a social sensor and utilize Kalman filtering and particle filtering for estimating the centers of earthquakes. Arase et al.[6] remarked that frequent trip patterns in geo-tagged photos during photo sharing on social media sites could be beneficial for tourism industries in that such information can assist tourism companies in recommending places of interest. Notably, such data may also allow tourism officials to conduct a review of the transportation system. Many different application domains now have the potential to use this geo-annotated user-generated data.

Novel text mining techniques are required for extracting and tracking local topics and events in geo-annotated documents. Traditional text mining techniques, however, have limitations when it comes to handling such data. Burstiness has been one of the most important criterion for extracting topics and events from documents posted on the Internet. The algorithm that has had the most significant influence on many studies is Kleinberg's burst

detection algorithm[3], which is based on a queuing theory for detecting bursty network traffic. Kleinberg's burst detection algorithm can be used for analyzing document streams from various sources such as e-mails[3], blogs[7], online publications[8], bulletin boards, and social tags [9].

Researchers have started to study the temporal bursts of local topics and events because of the growing number of georeferenced documents. Michael et al.[10] 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 Lappas et al.[11], for each region, a local document stream is located in "A", and a new method for finding temporal burstiness patterns among several regions was devised. A search engine that considers the spatiotemporal burstiness of terms in the process of document retrieval was demonstrated. Zimmermann et al.[12] presented a clustering method that detects, tracks, and updates large and small bursts of news in a two-level topic hierarchy. Both of these studies anticipated developing techniques for spatiotemporal burstiness. They did not address the need to identify location-based temporal bursts.

There are numerous studies involved with identifying and tracking local topics and events. Wang et al. and Yin et al.[13, 14] proposed new topic models to discover different topics in geographical regions. Furthermore, Yang et al.[15] developed a method to reveal the appearance and disappearance of topics in different regions. Canneyt et al.[16] developed a method that can detect new places of interest using Support Vector Machine (SVM) techniques to improve existing databases of places. In Lee et al.[17], a geo-social event detection method, which defines identities when local events are occurring in unusually crowded locations, is proposed. These techniques are effective for detecting local topics and events. They do not identify, however, the burstiness of local topics and events.

3. DETECTING BURSTS

In this section, a sequence of batched documents, bursts, and Kleinberg's burst detection algorithm are described

3.1 Sequence of Batched Documents

A sequence of batched documents is similar to a batched data stream. It is defined as a sequence of sets of documents arranged in order. Figure 1 illustrates a sequence of batched documents where the sets of documents, which are called batches, are posted in order. Suppose that there are n batches of documents $BDS = \{BD_1, BD_2, \ldots, BD_n\}$. Each batch of documents is a set of documents and is posted discretely. The time interval between

BD_t and BD_{t+1} has no meaning. An example of a sequence of batched documents is conference papers posted on publication sites. Moreover, tweets posted on the Twitter site at fixed time intervals (e.g., per hour, day, or month) are referred to as a sequence of batch documents.

3.2 Bursts

The number of documents that include some particular keywords related to a topic or an event increases sharply when the topic or event attracts the interest of many people. As the number of documents including the keywords becomes larger, the number of documents related to the topics or events in batches increases in a sequence of batched documents. The keywords are considered highly bursty during a distinct period in which the number of documents including the keywords is greater than usual. This bursty distinct period is referred to as a bursty period in a sequence of batched documents.

For example, tweets on the Twitter site are referred to as a sequence of document sets. If the topic "heavy rainfall" is attracting more attention from people, the number of documents including the keyword "rain" increases in the tweets, hour by hour, and eventually the number of tweets including the keyword "rain" becomes larger than usual. This distinct period can be extracted as a bursty period.

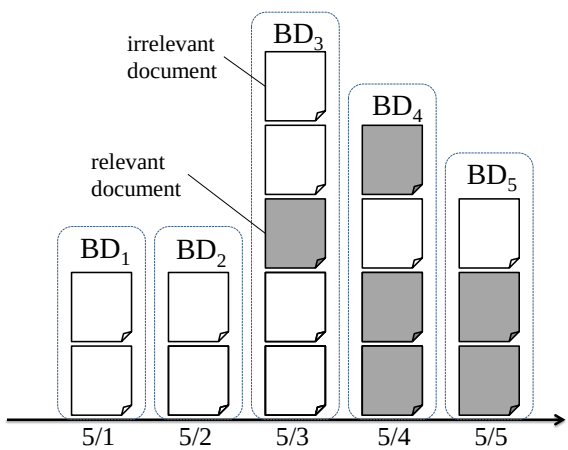


Figure 1. Example of a sequence of batched documents

3.3 Kleinberg's Burst detection Algorithm

Let BR_i be a set of relevant documents in BD_i. A relevant document includes a keyword that is an object of analyzing bursts. The number of documents in BD_i and BR_i are denoted by nbd_i and nbr_i, respectively. The sequence of the number of documents is $nbds = (nbd_1, nbd_2, \ldots, nbd_n)$, and

the sequence of the number of relevant documents is $nbrs = (nbr_1, nbr_2, \ldots, nbr_n)$. Let the total number of documents and relevant documents be denoted by:

$$NBD = \sum_{t=1}^{n} nbd_t \qquad (1)$$

$$NBR = \sum_{t=1}^{n} nbr_t \qquad (2)$$

For example, in Figure 1, there are five batched documents in the sequence; $nbds = (1, 2, 5, 4, 3)$, and $nbrs = (0, 0, 1, 3, 2)$.

Kleinberg defined a model with an infinite-state automaton where bursts are represented as state transitions. Assuming that there are m states in the automaton, the number of documents and the number of relevant documents are probabilistic outputs that depend on the internal states of the infinite-state automaton. The problem is defined as finding an optimal state-transition sequence $s = (s_1, s_2, \ldots, s_n)$ to minimize the following cost function:

$$C(s \mid nbds, nbrs) = \left(\sum_{i=1}^{n-1} \tau(s_i, s_{i+1}) \right) + \left(\sum_{i=1}^{n} -\ln \sigma_{s_i}(nbd_i, nbr_i) \right) \qquad (3)$$

The function $\tau(i,j)$ returns a state-transition cost from the i-th state to the j-th state. It is defined as:

$$\tau(i, j) = \begin{cases} (j - i)\gamma & if \quad j > i, \\ 0 & otherwise \end{cases} \qquad (4)$$

where γ (> 0) is a user-given parameter. Equation 4 indicates that moving to a higher state incurs a cost, and moving to a lower state incurs no cost. The function $\sigma_k(nbd_i, nbr_i)$ is the exponential density function for the probability of outputting nbd_i, and nbr_i in the k-th state, and is defined as:

$$\sigma_k(nbd_i, nbr_i) = \binom{nbd_i}{nbr_i} p_k (1 - p_k)^{(nbd_i - nbr_i)} \qquad (5)$$

where p_k is the arrival rate of documents associated with the k-th state. It is defined as:

$$p_k = \frac{NBR}{NBD}\beta^k, \quad (6)$$

where β (> 1.0) is a user-given parameter. Equation 6 indicates that a higher state has a higher rate of relevant documents.

The Viterbi algorithm for hidden Markov models, which is a dynamic programming approach, is the most effective solution for determining an optimal state-transition sequence:

$$C_j(i) = -\ln \sigma_k(nbd_i, nbr_i) + \min_l \left(C_l(i-1) + \tau(l, j)\right), \quad (7)$$

where $C_j(i)$ is the minimum cost of a state-transition sequence that ends with state j at the i-th batched document. Equation 7 can be calculated using the previous (i-1)-th $C_l(i\text{-}1)(1 \le l \le m)$. Then , we find the minimum cost in $C_l(n)$ $(1 \le l \le m)$. Suppose that the minimum cost in $C_l(n)$ $(1 \le l \le m)$ is $C_{min}(n)$ $(1 \le min \le m)$. Then , we trace back with $C_{min}(n)$ as the starting point.

4. PRELIMINARIES

This section presents some preliminaries, the data model, and the problem definition.

4.1 Data Model

In this study, a georeferenced document $gd_{i,j}$ consists of four items: date, text data, posted time, and location information: $gd_{i,j} = < date_{i,j}, text_{i,j}, pt_{i,j}, l_{i,j} >$, where $date_{i,j}$ is the date that the batched document was posted, $text_{i,j}$ is the text data (e.g., title, posted short message, and tags), $pt_{i,j}$ is the posted time, and $l_{i,j}$ is the location where the document was posted or is located (i.e., the latitude and longitude).

A sequence of batched georeferenced documents $SBGD$ = (BGD_1, BGD_2,…, BGD_n) is similar to a sequence of batched documents, where each of the batched documents represents a set of georeferenced documents. Let BGD_i be the i-th batched georeferenced document in $SBGD$: $BGD_i = \{gd_{i,1}, gd_{i,2}, \ldots, gd_{i,numd(i)}\}$, where $numd(i)$ is the number of georeferenced documents in the i-th batch of georeferenced documents. Figure 2 is an example of an $SBGD$ comprising five georeferenced document sets. The georeferenced document $gd_{3,3}$ is the 3rd georeferenced document in BGD_3, which is represented as a post on May 3rd and has a location in the geographical coordinate space.

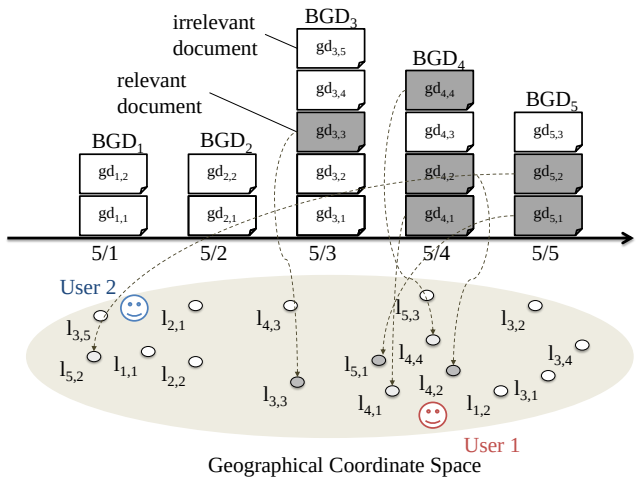


Figure 2. Example of a sequence of batched georeferenced documents

4.2 Problem Definition

Let BRG_i be a set of relevant georeferenced documents, where each georeferenced document in it includes a keyword that is an object of detected bursts. Each relevant georeferenced document's text data includes the keyword *key*: $BRG_i = \{rg_{i,1}, rg_{i,2}, \ldots, rg_{i,numd(i)}\}$, where $numd(i)$ is the number of documents in the i-th set of the relevant georeferenced documents. A set of relevant georeferenced documents BRG_i is a subset of BGD_i:

$$\varphi_i : BRG_i \rightarrow BGD^+, \qquad rg_{i,j} \mapsto gd_{i,\phi_i(j)}, \tag{8}$$

In Figure 2, there are four georeferenced documents in BRG_4, and there are three georeferenced relevant documents in BRG_4. In this example, $rg_{4,1} = gd_{4,1}$, $rg_{4,2} = gd_{4,2}$, and $rg_{4,3} = gd_{4,4}$. Therefore, $\varphi_i(1) = 1$, $\varphi_i(2) = 2$, and $\varphi_i(3) = 4$.

The sequence of the number of batched georeferenced documents is $nbgds = (nbgd_1, nbgd_2, \ldots, nbgd_n)$, and the sequence of the number of relevant georeferenced documents is $nbrgs = (nbrg_1, nbrg_2, \ldots, nbrg_n)$. Let the total number of georeferenced documents and relevant georeferenced documents be denoted by:

$$NBGD = \sum_{t=1}^{n} nbgd_t \tag{9}$$

$$NBRG = \sum_{t=1}^{n} nbrg_t \qquad (10)$$

The objective of this study is to identify the local burstiness of the keyword *key* including a local topic or event considering the user's location. This is defined as a problem that finds an optimal state-transition sequence $s = (s_1, s_2, \ldots, s_n)$ minimizing the following cost function. The temporal period is extracted using the optimal state-transition sequence:

$$C(s \mid nbgds, nbrgs) = \left(\sum_{i=1}^{n-1} \tau(s_i, s_{i+1}) \right) + \left(\sum_{i=1}^{n} -\ln \hat{\sigma}_{s_i}(nbgd_i, nbrg_i) \right) \qquad (11)$$

where

$$\hat{\sigma}_k(nbgd_i, nbrg_i) = \binom{nbgd_i}{nbrg_i} \tilde{p}_k (1 - \tilde{p}_k)^{(nbgd_i - nbrg_i)} \qquad (12)$$

$$\tilde{p}_k = \frac{NBRG}{NBGD} \beta^k \qquad (13)$$

For example, in Figure 2, *User*1 is located close to the location of the georeferenced relevant documents that are in bursty discrete periods. For *User*1, we must indicate that *key* is highly bursty. Conversely, *User*2 is distant from the georeferenced relevant documents that are in bursty discrete periods. For *User*2, we show that *key* is not highly bursty.

5. PROPOSED ALGORITHM

In this section, we propose a novel algorithm that extracts location-based bursts from a sequence of batched georeferenced documents.

5.1 Main Concept

The most basic method to extract location-based bursts is to use a cutoff distance. If a user wishes to know the local burstiness of *key* nearby, only georeferenced documents within *cutoff* from the user are selected by the algorithm. In this algorithm, we find an optimal state-transition sequence to minimize $C(s|nbgds, nbrgs)$, where *nbgds* and *nbrgs* are the sequences of georeferenced documents that satisfy $dist(l_{i,j}, ul) < cutoff$ and relevant geographical documents that satisfy $dist(l_i, \phi_{i(j)}, ul) < cutoff$.

The advantage of this algorithm is its simplicity. A disadvantage is that it is highly dependent on the cutoff distance. For example, if the cutoff distance *cutoff* is small for a user, the method recognizes that *key* is not

highly bursty. This issue can be avoided by setting a large value for the cutoff distance *cutoff*. This, then, results in another issue: burst detections are visibly affected by georeferenced documents distant from the user.

To identify the local burstiness seamlessly, we have integrated the influence rates of georeferenced documents, which are determined by the distance from the user in Kleinberg's enumerating burst detection algorithm. In our previous work, the forgetting theory was used to calculate the influence rate. Georeferenced documents gradually lose their weight (or memory) according to an increase in distance from the user.

5.2 Influence Rate

In our previous work, the influence rates of the georeferenced documents were determined by their distance from the user. This method did not consider the relationships between georeferenced documents in location-based burst detection. To overcome this issue, we integrate density-based criteria into the location-based burst detection algorithm to identify the local burstiness more precisely. In this study, we utilize kernel density estimation to estimate the densities of the georeferenced documents.

Kernel density estimation can estimate the overall distribution from the location information dataset. In this study, the following kernel function is used:

$$den\,(gd_{i,j}) = \sum_{k=1}^{n} \sum_{l=1, dist\,(gd_{k,l}, gd_{i,j}) \le \varepsilon}^{numd\,(k)} \left(\frac{3}{\pi\varepsilon^2} \left(1 - \frac{dist\,(gd_{k,l}, gd_{i,j})}{\varepsilon^2} \right) \right) \quad (14)$$

where ε is the kernel bandwidth that is used to limit the calculation of the distance between georeferenced documents, and the function *dist* returns the distance between two input georeferenced documents. Although we can estimate the local densities of georeferenced documents using kernel density estimation, kernel density estimation does not consider temporal changes in density. To consider temporal changes, the forgetting theory is used to decrease influence:

$$den\,(gd_{i,j}) = \sum_{\substack{k=1 \\ l=1}}^{n} \sum_{\substack{dist\,(gd_{k,l}, gd_{i,j}) \le \varepsilon, \\ alt\,(gd_{k,l}, gd_{i,j}) \le \delta}}^{numd\,(k)} \frac{3}{\pi\varepsilon^2} \left(1 - \frac{dist\,(gd_{k,l}, gd_{i,j})}{\varepsilon^2} \right) \times \xi^{\,alt\,(gd_{k,l}, gd_{i,j})} \quad (15)$$

where δ is the time bandwidth that can limit the calculation of the inter-arrival time between georeferenced documents: ($0 < \xi < 1$) is a user-given parameter, and the function *alt* returns the time interval between two input georeferenced documents. Equation 15 indicates that two

georeferenced documents posted at a similar time have a strong relationship. Otherwise, they have a weak relationship. The degree of relationship decreases gradually according to the forgetting factor ξ.

The definition of the influence rate of a georeferenced document $gd_{i,j}$ is defined as:

$$infr_{i,j} = \frac{den(gd_{i,j}) - den_{\min}}{den_{\max} - den_{\min}} \times \alpha^{dist(gd_{i,j}, ul)} \quad (16)$$

where $\alpha(0 < \alpha < 1)$ is a user-given parameter, den_{max} and den_{min} are the maximum value and the minimum value of the densities of georeferenced documents, respectively.

5.3 Algorithm

Let the set of the influence rates in BGD_i be $INFR_i = \{infr_{i,1}, infr_{i,2}, \ldots, infr_{i,numd(i)}\}$. In the proposed location-based burst detection algorithm, we replace *nbgds* and *nbrgs* by the total amount of influence rates of the georeferenced documents. The new definitions of the number of georeferenced documents and the number of relevant georeferenced documents are as follows:

$$nbgd_i = \sum_{j=1}^{numd(i)} infr_{i,j} \quad (17)$$

$$nbrg_i = \sum_{j=1}^{numd(i)} infr_{i,\phi_i(j)} \quad (18)$$

The steps of the proposed location-based burst detection algorithm are as follows:

1. For each georeferenced document, the density of the georeferenced document is calculated.
2. For each batched georeferenced document, the set of the influence rates $INFR_i$ for BGD_i is created by calculating the influence rates of the georeferenced documents. The influence rate of each georeferenced document is calculated in accordance with the location of user *ul* and its local density.
3. The elements $nbgd_i$ and $nbrg_i$ are calculated using $INFR_i$.
4. The sequences *nbgds* and *nbrgs* are created using the elements $nbgd_i$ and $nbrg_i$, respectively.

5. The optimal state-transition sequence $s = (s_1, s_2, \ldots, s_n)$ is found to minimize Equation 11, where the input sequences are the sequences *nbgds* and *nbrgs*.

6. EVALUATION

To evaluate the location-based burst detection algorithm that considers the user's location, we used an actual *SBGD* that was constructed by crawling geo-tagged tweets posted on Twitter. We collected 480,000 geo-tagged tweets from Twitter using its API. The period for the set of tweets was December 2011 to February 2012 (JST). In the experiments, we used the keyword "snow," which achieved the highest score in the tf*idf results. The five major cities of Japan, Fukuoka, Hiroshima, Osaka, Nagoya, Tokyo, and Sendai were set as the users' positions (Figure 3). The parameters used in the experiments were $\varepsilon = 0.5$, $\delta = 5$, $\xi = 0.9$, and $\alpha = 0.9$. The parameters for Kleinberg's burst detection algorithm used in the experiments were $\beta = 1.1$, and $\gamma = 0.01$.

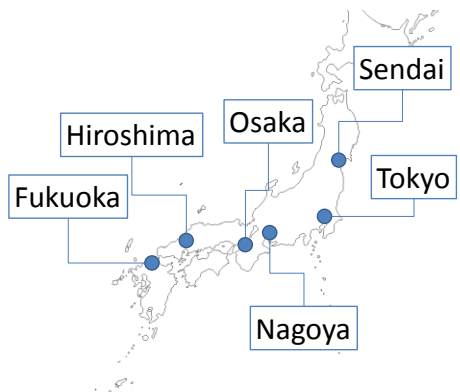


Figure 3. Map of Japan

The graphs in Figure 4 present the results for Fukuoka, Hiroshima, Osaka, Nagoya, Tokyo, and Sendai. They depict the burstiness from January 1 to January 31 (JST). The conventional location-based burst detection algorithm[4] is denoted by LBD, and the proposed location-based burst detection algorithm is denoted by LBDWD. LBD is our previous algorithm, which is a location-based burst detection algorithm based on Kleinberg's burst detection algorithm that identifies bursty phenomena in a sequence of batched georeferenced documents. These results indicate that location-based bursts can be detected. We verified that each burst result was correct. In early January, the area of western Japan had snow, therefore, bursts occurred in Fukuoka, Hiroshima, and Osaka (Figures 4 (a), (b), and (c)).

However, bursts did not appear in Nagoya and Tokyo (Figures 4 (d) and (e)) because they are located in eastern Japan.

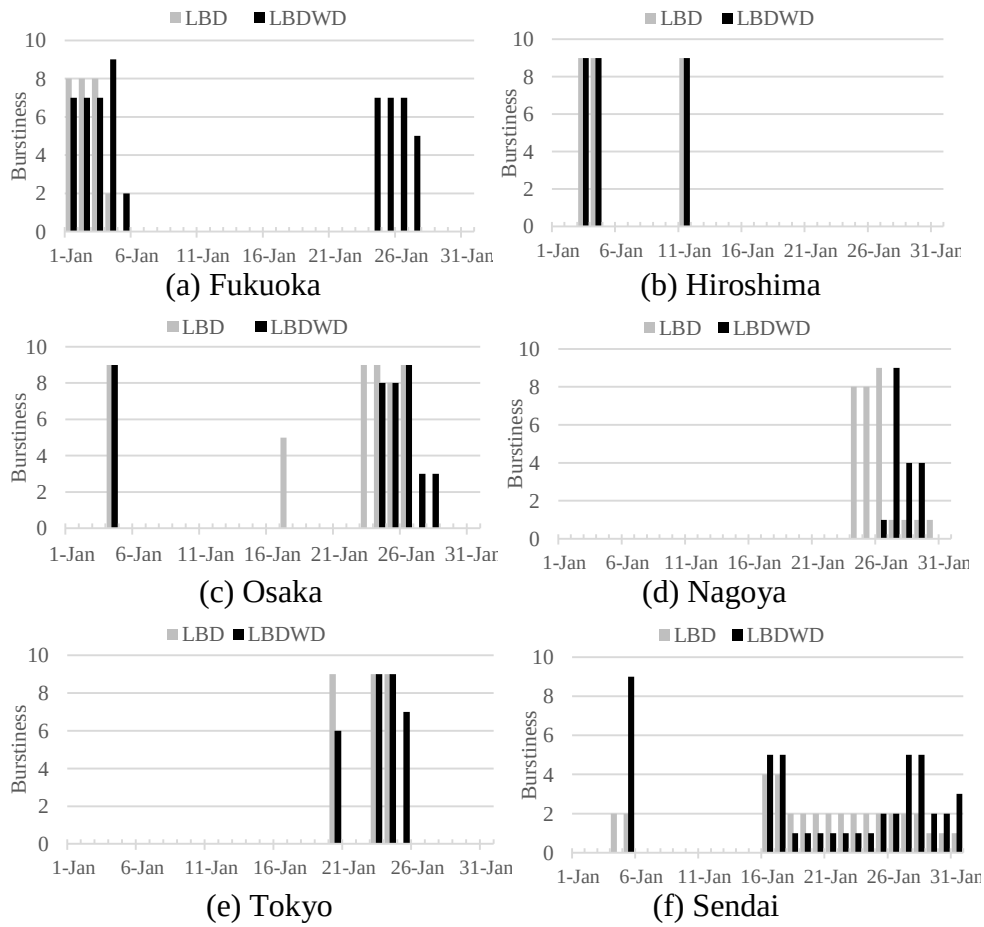


Figure 4. Local Bursts of Fukuoka, Hiroshima, Osaka, Nagoya, Tokyo, and Sendai (these graphs depict the burstiness from January 1 to January 31 (JST))

The graphs in Figure 5 present the results for Fukuoka, Hiroshima, Osaka, Nagoya, Tokyo, and Sendai. These graphs depict the burstiness from February 1 to February 21 (JST). In early February, three bursts occur in Fukuoka when we used LBD. However, these were noise bursts (Figure 5(a)). From February 18 to 19, it snowed in Fukuoka. Both of these event location-based bursts were detected correctly. In Hiroshima (Figure 5(b)), from February 1 to 3, bursts occur. These bursts were noise because of the small number of target tweets. Figures 5(c) and (d) present the results of Osaka and Nagoya, respectively. In early February, the Kinki and Tubu regions, where Osaka and Nagoya are located, experienced snow. Therefore,

these bursts occurred. In Tokyo (Figure 5(e)), bursts occurred from February 16 to February 18. In the Tokyo metropolitan region, there was heavy snow, therefore these bursts are correct.

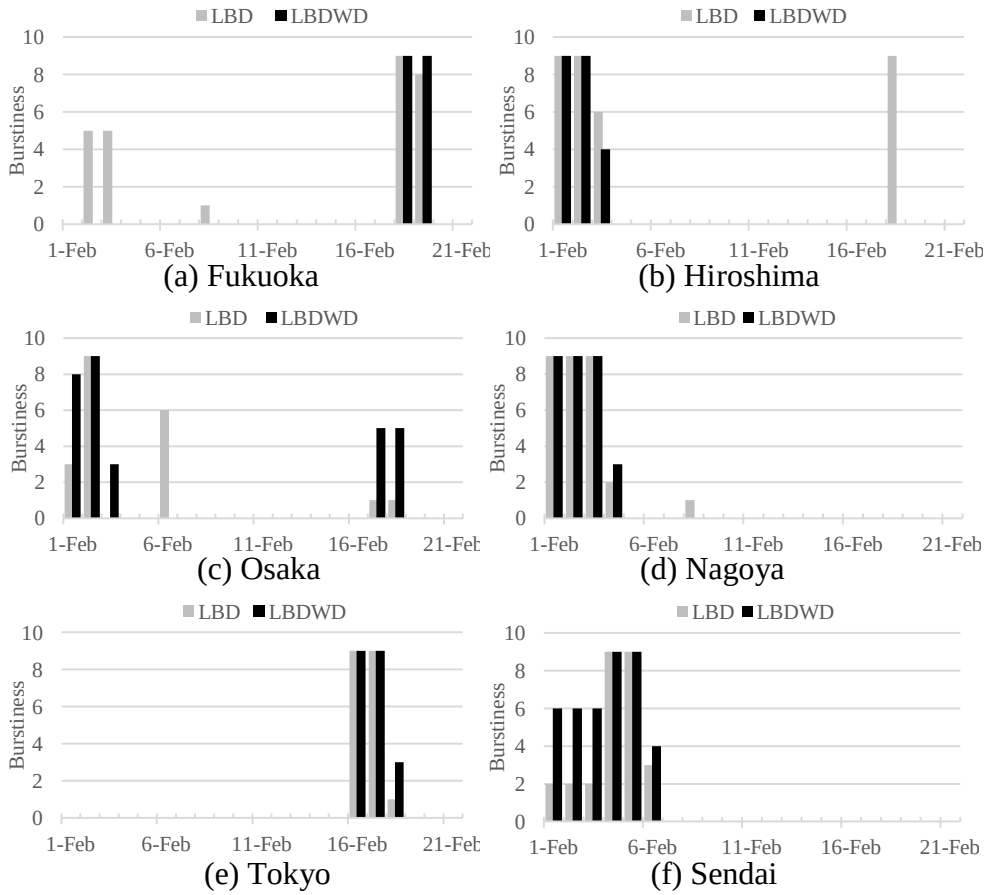
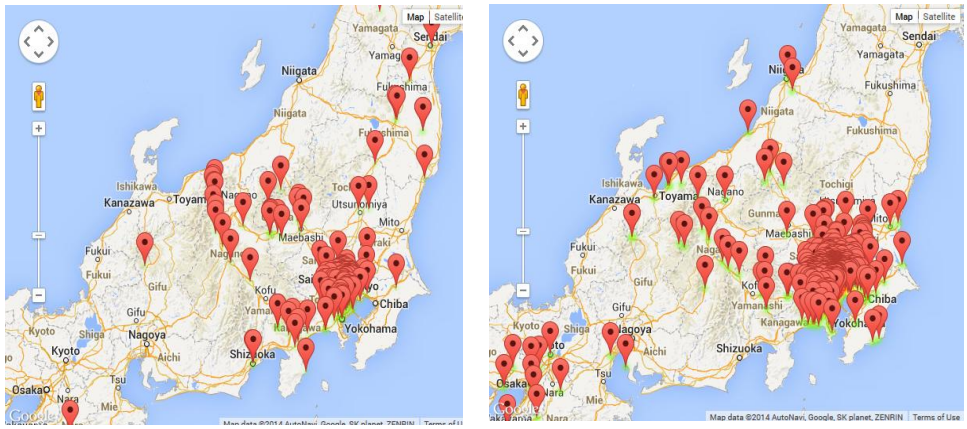


Figure 5. Local Bursts of Fukuoka, Hiroshima, Osaka, Nagoya, Tokyo, and Sendai (these graph depict the burstiness from February 1 to February 21 (JST))

To verify the detected location-based bursts, we verified the distribution of the tweets on the map. Figures 6(a) and (b) illustrate the distribution of the tweets on January 20 and January 23, respectively. Bursts did not occur in Nagoya and Osaka because these cities are distant from Tokyo. These figures indicate that we can detect location-based bursts correctly.

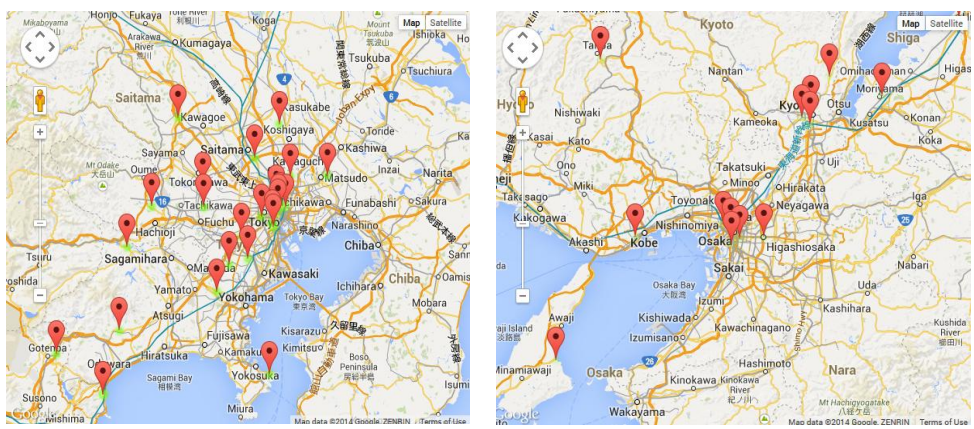
Moreover, we compared LBD and LBDWD. Figure 7 is the distribution of the tweets in Tokyo and Osaka. Figure 4(e) indicates that a burst occurred on January 25 in Tokyo when we used LBDWD. Figure 7(a)

displays the distribution of the tweets on the map. The density of the tweets is high, therefore LBDWD can detect location-based bursts correctly. Figure 5(c) indicates that a burst occurred on February 8 in Tokyo when we used LBD. Figure 7(b) shows the distribution of the tweets on the map. The density of the tweets is not high in Osaka. Therefore, this burst is in error. This result indicates that LBDWD can detect location-based bursts more accurately than LBD.



(a) January 20 (b) January 23

Figure 6. Distribution of tweets including the keyword “snow” in Tokyo



(a) Tokyo on Jan 25 (b) Osaka on Feb 8

Figure 7. Distribution of tweets including the keyword “snow” in Tokyo and Osaka

7. CONCLUSION

In this paper, we proposed a novel location-based burst detection algorithm for identifying local burstiness in a sequence of batched

georeferenced documents composed of ordered georeferenced document sets. Burstiness is one of the simplest yet most robust criterion for identifying trending topics and events in a sequence of batched documents. Identifying local burstiness allows us to extract not only the peak periods, but also the locality of trending topics and events. The proposed location-based burst detection algorithm utilizes kernel density estimation to integrate the relationships between georeferenced documents in location-based burst detection. To evaluate the proposed location-based burst detection algorithm, we used an actual sequence of batched georeferenced documents constructed by crawling tweets posted on the Twitter site. The experimental results confirm that the proposed location-based burst detection algorithm could identify the burstiness of a keyword related to a weather topic in Japan more precisely compared with the algorithm proposed in our previous work. In a future work, we will develop an on-line algorithm that will detect location-based bursts in real-time.

8. ACKNOWLEDGMENT

This work was supported by a Hiroshima City University Grant for Special Academic Research (General Studies) and JSPS KAKENHI Grant Number 26330139.

REFERENCES

[1] M. Naaman, Geographic information from georeferenced social media data. *SIGSPATIAL Special*, 3(2), p54-61, 2011. http://dx.doi.org/10.1145/2047296.2047308.

[2] B.J. Hecht, and D. Gergle, On the "Localness" of user-generated content. In K. Inkpen, C. Gutwin, and J. Tang (Eds.), *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW)* (p229-232). Savannah, GA, USA: ACM Press, 2010. http://dx.doi.org/10.1145/1718918.1718962.

[3] J. Kleinberg, Bursty and hierarchical structure in streams. In O.R. Zaianne, R. Goebel, D. Hand, D. Keim, and R. Ng (Eds.), *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (p91-101). Edmonton, AB, Canada: ACM Press, 2002. http://dx.doi.org/10.1023/A:1024940629314.

[4] K. Tamura, and H. Kitakami, Detecting location-based enumerating bursts in georeferenced micro-posts. In T. Matsuo, K Hashimoto, and S Hirokawa (Eds.), *Proceedings of the 2013 Second IIAI International Conference on Advanced Applied Informatics (IIAI-AAI)* (p389-394). Matsue, Japan: IEEE Computer Society, 2013. http://dx.doi.org/10.1109/IIAI-AAI.2013.36.

[5] T. Sakaki, M. Okazaki, and Y. Matsuo, Earthquake shakes twitter users: Real-time event detection by social sensors. In M. Rappa, P. Jones, J. Freire, and S. Chakrabarti (Eds.), *Proceedings of the 19th International Conference on World Wide Web (WWW)* (p851-860). Raleigh, NC, USA: ACM Press, 2010. http://dx.doi.org/10.1145/1772690.1772777.

[6] Y. Arase, X. Xie, T. Hara, and S. Nishio, Mining people's trips from large scale geo-tagged photos. In A.D. Bimbo, S. Chang, and A. Smeulders (Eds.), *Proceedings of the eighteenth ACM International Conference on Multimedia (MM)* (p133-142). Firenze, Italy: ACM Press, 2010. http://dx.doi.org/10.1145/1873951.1873971.

[7] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, On the bursty evolution of blogspace. In G. Hencsey, B. White, Y.R. Chen, L. Kovacs, and S. Lawrence (Eds.), *Proceedings of the 12th International Conference on World Wide Web (WWW)* (p568-576). Budapest, Hungary: ACM Press, 2003. http://dx.doi.org/10.1007/s11280-004-4872-4.

[8] K.K. Mane, and K. Borner, Mapping topics and topic bursts in PNAS. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1), p5287-5290, 2004. http://dx.doi.org/10.1073/pnas.0307626100.

[9] J. Yao, B. Cui, Y. Huang, and X. Jin, Temporal and social context-based burst detection from folksonomies. *Paper Presented at the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI)* Atlanta, Georgia, USA, July 11-15, 2010.

[10] M. Mathioudakis, N. Bansal, and N. Koudas, Identifying, attributing and describing spatial bursts. *Proceedings of the VLDB Endowment*, 3(1-2), p1091-1102, 2010. http://dx.doi.org/10.14778/1920841.1920978.

[11] T. Lappas, M.R. Vieira, D. Gunopulos, and V.J. Tsotras, On the spatiotemporal burstiness of terms. *Proceedings of the VLDB Endowment*, 5(9), p836-847, 2012. http://dx.doi.org/10.14778/2311906.2311911.

[12] M. Zimmermann, I. Ntoutsi, Z.F. Siddiqui, M. Spiliopoulou, and H.-P. Kriegel, Discovering global and local bursts in a stream of news. In S. Ossowski and P. Lecca (Eds.), *Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC)* (p807-812). Trento, Italy: ACM Press , 2012. http://dx.doi.org/10.1145/2245276.2245433.

[13] C. Wang, J. Wang, X. Xie, and W.-Y. Ma, Mining geographic knowledge using location aware topic model. In R. Purves and C. Jones (Eds.), *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval (GIR)* (p65-70). Lisbon, Portugal: ACM Press, 2007. http://dx.doi.org/10.1145/1316948.1316967.

[14] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang, Geographical topic

discovery and comparison. In S. Sadagopan, K. Ramamritham, A. Kumar, M.P. Ravindra, E. Bertino, and R. Kumar (Eds.), *Proceedings of the 20th International Conference on World Wide Web (WWW)*, (p247-256). Hyderabad, India: ACM Press, 2011. http://dx.doi.org/10.1145/1963405.1963443.

[15] H. Yang, S. Chen, M.R. Lyu, and I. King, Location-based topic evolution. In S.H. Gary Chan, E.Y. Chang, and M. Lyu (Eds.), *Proceedings of the 1st International Workshop on Mobile Location-based Service* (p89-98). New York: ACM Press, 2011. http://dx.doi.org/10.1145/2025876.2025894.

[16] S. Van Canneyt, O. Van Laere, S. Schockaert, and B. Dhoedt, Using social media to find places of interest: a case study. In D. Pfoser and A. Voisard (Eds.), *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information (GEOCROWD)* (p2-8). New York: ACM Press, 2012. http://dx.doi.org/10.1145/2442952.2442954.

[17] R. Lee, S. Wakamiya, and K. Sumiya, Discovery of unusual regional social activities using geo-tagged microblogs. *World Wide Web*, 14(4), p321-349, 2011. http://dx.doi.org/10.1007/s11280-011-0120-x.