Expert Systems with Applications 39 (2012) 9623–9641

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Mining spatio-temporal information on microblogging streams using a density-based online clustering method

Chung-Hong Lee

Department of Electrical Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan

ARTICLE INFO

Keywords:
Topic detection
Text mining
Microblogging
Temporal analysis
Spatial analysis

ABSTRACT

Social networks have been regarded as a timely and cost-effective source of spatio-temporal information for many fields of application. However, while some research groups have successfully developed topic detection methods from the text streams for a while, and even some popular microblogging services such as Twitter did provide information of top trending topics for selection, it is still unable to fully support users for picking up all of the real-time event topics with a comprehensive spatio-temporal viewpoint to satisfy their information needs. This paper aims to investigate how microblogging social networks (i.e. Twitter) can be used as a reliable information source of emerging events by extracting their spatio-temporal features from the messages to enhance event awareness. In this work, we applied a density-based online clustering method for mining microblogging text streams, in order to obtain temporal and geospatial features of real-world events. By analyzing the events detected by our system, the temporal and spatial impacts of the emerging events can be estimated, for achieving the goals of situational awareness and risk management.

1. Introduction

1.1. Motivations

Recently a growing number of internet users keep up with newest information by utilizing social media tools (e.g. Twitter), searching for a hot news topic about some emerging events. However, while some research groups have successfully developed topic detection methods from the text streams for a while, and even some popular microblogging services like Twitter did provide information of top trending topics for selection, it is still unable to fully support user for picking up all of the real-time event topics with a comprehensive spatio-temporal viewpoint to satisfy their information needs.

In this work, we take *Twitter* as our microblogging data source to identify the problem domain, for describing and realizing evolving events in depth. In this work, an "*event*" is taken to be something that happens at some specific time and place (e.g. *an earthquake striking Japanese cities on March 2011*). Twitter, regarded as one of powerful social media tools, allows users to post short messages (i.e. maximum 140 characters, also known as "tweet") to communicate each other. In particular, once people suddenly suffer from some unexpected disasters, over thousands of Twitter users seized their mobile phones or computers to peck out tweets for communication in the form of text messages, web-based instant messages, or posts on Twitter's site. The vest amount of tweets might cause Twitter's datasets to be fairly difficult to process. Hence, the application platforms should be able to process large volumes of such textual data arriving over time in a stream.

On the other hand, severe natural disasters such as earthquakes, tsunami, etc., require new scientific methodologies for early warning of risk and real-time event awareness. Understanding their possible impacts and striving towards their timely detection and prevention can help protect lives and properties. Under such circumstances, the location from which a tweet was issued can be an enormous help. Messages from mobile phones with GPS receivers can contain location information. Recently Twitter also allowed such information to be attached to tweets, in order to enable people to apply this data to produce more relevant location based real-time results. Unfortunately, enabling the functions of latitude and longitude features on tweets is optional, up to users' choices. Therefore, in this work we developed an online spatio-temporal information platform which can generate really useful results that are obviously impactful in real-world applications.

1.2. Problem statements

Given the fact that some perspectives of on-going and unknown events are still better observed by human eyes and, in some cases news reporters are unable to get the event information, messages in tweets about the events could prove to be useful. For example, in many cases, when some earthquake event occurred, a few

E-mail address: leechung@mail.ee.kuas.edu.tw

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.02.136

seconds and minutes after the quake, people eager to talk to their families by phones or other communication tools, and often they found the only live media for realizing the situation related to the new quake were – **tweets**. Under such a circumstance, an actual demand is to develop a way for discovering real-time event truth with temporal and geospatial information from microblogging messages to offer users insightful information in an efficient manner.

Recently Twitter released some part of location service that enables mobile users publish their tweets with geospatial data such as latitude and longitude. Such services encourage more research work involved in mining the spatio-temporal information on microblogs to get real-time and geospatial event information. Mining the spatio-temporal information related to critical events is a challenging task, which attempts to extract useful information from large volume of continuously arriving tweet streams. Hence, in this work we are keen to explore the potential of spatio-temporal information provided by Twitter messages, thus contributing to a satisfaction of the information need of 'situational awareness' for event control.

1.3. Research objectives

In this work, we study the problem of event detection and awareness by monitoring on-line real-time messages (i.e. Twitter messages), and exploiting the geolocation data provided by the experimental social networking datasets for situational awareness. To our best knowledge, previous researches those try to extract hot topics from Twitter in real time focused on only temporal models. This research also extracts spatial information of each topic. This is a novel approach in this research area. Alone with the event detected from the Twitter messages, we attempt to reverse engineer the location of a tweeted event from text analysis because some location information cannot be acquired for free public Twitter data. The aim of this work is to study how microblogging social networks (i.e. Twitter) can be used as a reliable information source of emerging events by extracting their spatio-temporal features from the messages to enhance event awareness. We believe that Twitter could be used to detect events and notify users who are concerning event development, by applying Twitter information to establish a spatio-temporal model for event estimation, which is able to find the center and trajectory of event location. Hence, in this work we developed several algorithms for mining Twitter text streams to obtain real-time and geospatial event information. The significance of the work lies more in the application than in the modeling algorithms. The proposed solution is being described in the following sections.

2. System framework and approaches

In this chapter, we describe our system framework and approaches. First of all, we present some problem characteristics and difficulties in system development for acquiring spatio-temporal information associated with the discovered events as below.

- There are hundreds of thousands Twitter messages and spams performing data exchange on the internet incessantly. The source of Twitter messages is an open-ended data stream and the amount of the accumulated data is extremely large, so it is impossible to allow all data be loaded into the memory for computation. Thus, an effective incremental learning approach is essentially required for discovering knowledge from such text streams. In general, there are two fundamental data mining techniques that can be considered in conjunction with Twitter data: (a) graph mining using analysis of the links among messages, and (b) text mining based on analysis of the messages' textual contents. In this work, we mainly focus on investigation of text mining methods for mining Twitter data.
- Once dealing with Twitter streams, one important indication of change is the presence of ***bursts***. A burst in the Twitter messages implies that the occurrence of a certain topic feature is unexpectedly frequent in a short period of time. Such situations normally indicate some real-world event has now drawing much attention by Twitter users. The burst detection method applied in our work is concerned with automatic identification of bursts from Twitter posted messages, providing useful insights into the unusual events and in turn facilitating timely event monitoring.
- Real-time operation is a critical factor for the use of Twitter messages. User posts a tweet with some specific timestamp which indicates what someone says has happened at the specific time point and places. It is believed that the messages act as a useful lens into the social perception of an event in any region, at any point in time. For instance, people posted their tweets when earthquake occurred, the importance of these tweets are valuable just at the moment. To cope with temporal dynamics of tweets, in this work we developed a dynamic weighting scheme called *BursT* to adapt to such requirements, which is able to subtly reflect the changes over time and quickly assign proper weights for achieving accurate temporal evaluation of messages in such a dynamic environment.
- According to recent investigation of most users' location entries on Twitter, lots of users did not provide real location information, often incorporating fake location that can mislead most geographic event detection systems. Furthermore, even the tweet samples with the data of Twitter's *geo-tagging* location service that enables mobile users publish their tweets with latitude and longitude data, it is still not precise enough for detecting accurate location associated with the occurrence of some specific event only by utilizing such information. This is due to the fact that lots of tweets were sent by mobile devices and, the mobility of Twitter users make it difficult to detect actual fixed location related to the ongoing events.
- A difficult problem alone with online learning on the continuously incoming text streams is that the concept of interest behind content may change with time, depending on some *hidden context*, which is not given explicitly in the form of initial features. The phenomenon is known as *concept drift*. In the case of microblogs, this happens to produce a topic drift on messages. For microblogging services, it is obvious that the hot topics of some issue discussed by Twitter users often drift with time, depending on the newest development of the original event and some hidden context. For instance, in the case of Fukushima nuclear accidents, the hot topics on tweets starts with "earthquake" and "tsunami", and then move to "nuclear and radiation accident", and "supply chain risk", etc. A challenge in handling concept drift is distinguishing between true concept drift and noise, since some algorithms may overreact to noise, mistakenly interpreting it as concept drift.
- The topic transition related to evolving events is generally hard to be detected. If we have background knowledge such as "earthquake → tsunami → nukes", we may be able to understand the obtained clustering result. But it would be quite difficult to analyze transitions of unknown events. Extraction of meaningful information from a clustering result is not an easy task.

To solve the aforementioned issues, we have proposed a framework for event detection by spatio-temporal information discovered from Twitter messages. Our proposed system framework mainly consists of two modules, say, *content and temporal analysis*

module and *spatial analysis* module. The *content and temporal analysis* module is developed for handling microblogging message streams, and categorized them into thematic topics. Subsequently, the module of *spatial analysis* performs allocating topic centric messages to appropriate locations in the real world map. The system framework is developed based on *a density-based online clustering method*. As mentioned above, the concepts that we attempt to learn from the text streams may drift with time. To flexibly react to concepts drift in the messages, we have developed algorithms in the density-based clustering method using the *sliding window* technique, which are able to detect context changes without being explicitly informed about them. Detailed description of the technique is being discussed in Sections 3 and 4.

2.1. Assumptions and system framework

Prior to our discussion on our system architecture for mining spatio-temporal event information, we present several assumptions made for system development.

Assumption 1. The gathering messages tend to be a phenomenon of "temporal locality".

In this work, "event" is regarded as a set of messages that are highly concentrated on some issues in a period of time. Such a phenomenon is also described as the characteristics of temporal locality among messages. The concept of temporal locality is used to present that an event that is discussed at one point in time will be discussed again sometime in the near future.

Assumption 2. The messages associated with a real-world event are of a nature of event lifecycle.

To process incoming texts with a chronological order, a fundamental issue we concerned is how to find the significant features in text streams. In classic text retrieval systems, the most common method for feature extraction is to deal with each document as a bag-of-words representation. Such an approach is not completely suitable for our dynamic system. The main technical issue of detecting events in text streams is to derive a set of features (words) to describe each message and a similarity measure between messages (Lee, Wu, & Chien, 2011). It has been observed that, in microblogging text streams, some words are "born" when they appear the first time, and then their intensity "grow" in a period of time till reach a peak. These words are called *burst words*. As time passes by, once the topics are no longer discussed by people, they "fade away" with power law and eventually the feature words become "death" (disappear), or change to a normal state. Such a phenomenon is regarded as a lifecycle of the selected features associated with a particular event under investigation.

Assumption 3. Most messages tend to be a phenomenon of "spatial locality".

For some events, particularly the events related to some disasters (e.g. flooding), the affected areas were often being significantly migrated or expanded as time passes by. Initially the popular event topic largely represent the common interests of local users, and then once the event spreads, people in other places start to concern about it. *Spatial locality* is described as a set of messages concerning some topic, which are highly densely located in a specific geographical area. Such a phenomenon is considered in our work to derive algorithms for estimating location of an event by using location feature vectors. These concepts will be further described in Section 3.2.

Assumption 4. Most related locations of events can be obtained by content-based learning methods.

Ideally, as mentioned previously, the most intuiting method to obtain spatial information of messages is to annotate them with geographic coordinates which are based on a precise form of location (i.e. latitude and longitude). Unfortunately, enabling the functions of latitude and longitude features on tweets is optional, up to users' choices. Due to the consideration of privacy, the location information contained in a tweet structure is quite limited; only the time-zone geographic information on tweets can be obtained from user profiles. On the other hand, owing to the fact that some tweets are posted by the mobile devices (e.g. smart mobile phone), the event location and the location information extracted from tweet tags may not be exactly the same place. Under such a circumstance, we argue that, the strategy on tracking event location only by means of the feature of geographic coordinates on tweets is not essentially required, which may possibly mislead the results of the related studies. As a result, extracting location information (i.e. keywords) from text content seems to be a sensible solution. Although the methods of event tracking by monitoring specific strings might loss some hidden information due to some user-generated contents and critical messages may not contain any location keywords, in our system framework we develop effective mining approach to overcome such a limitation.

Accordingly, we develop a novel spatio-temporal topic detection system framework, as illustrated in Fig. 1. The system is designed as a two-pass process, consisted of two modules, say, *content and temporal analysis* module and *spatial analysis* module. As shown in Fig. 1, the *content and temporal analysis* module is developed for handling Twitter streams, and categorized them into thematic topics, which has been reported in our previous work (Lee, Chien, & Yang, 2010; Lee et al., 2011). Subsequently, the module of *spatial analysis* performs allocating topic centric messages to appropriate locations in the real world map Fig. 2 illustrates an overview of proposed microblogging topic detection system.

2.2. Content and temporal analysis (1st pass)

In order to effectively detect emerging events, our work started with mining hot topic news from numerous contextual posts on microblogging messages. In this work, "event" is regarded as a set of messages that are highly concentrated on some issues in a period of time. Such a phenomenon is also described as the characteristics of *temporal locality* among messages. Suppose that we have a temporally-ordered message stream $M = \{m_1, m_2, \ldots m_k, m_{k+1}, \ldots\}$, which arrived at time $T = \{t_1, t_2, \ldots t_k, t_{k+1}, \ldots\}$, $\forall m_k, m_j \in M$, $\nexists\ m_k = m_j$. First, a language filter will filter out some incoming messages which contain non-ASCII characters (i.e. Chinese, Japanese, etc.), and then the system decompose the textual messages into a bag-of-words feature. Subsequently, our algorithm started with construction of a dynamic feature space which maintains messages with *sliding window model* to deal with the message streams. New incoming messages will be reserved in memory till they are out of the time window. This process model can prevent the memory limitation problem caused by continuously coming messages. Then we utilized a *dynamic term weighting scheme* (Lee et al., 2011) to assign dynamic weights to each word, by comparing historical records. The *neighborhood generation algorithm* is performed to quickly establish relations with messages, and carry out the operation of *text stream clustering*. In this work, we utilized *IncrementalDBSCAN* as our online clustering algorithm. Therefore, the system constantly groups messages into topics, and the shape of clusters would change over time. Finally, hot topic events on microblogs can be determined and ranked by analyzing the collected cluster records.

The Twitter messages continuously posted by users around the world; so it is almost impossible to store all messages at one time due to the restrictions of memory limitation and constantly time

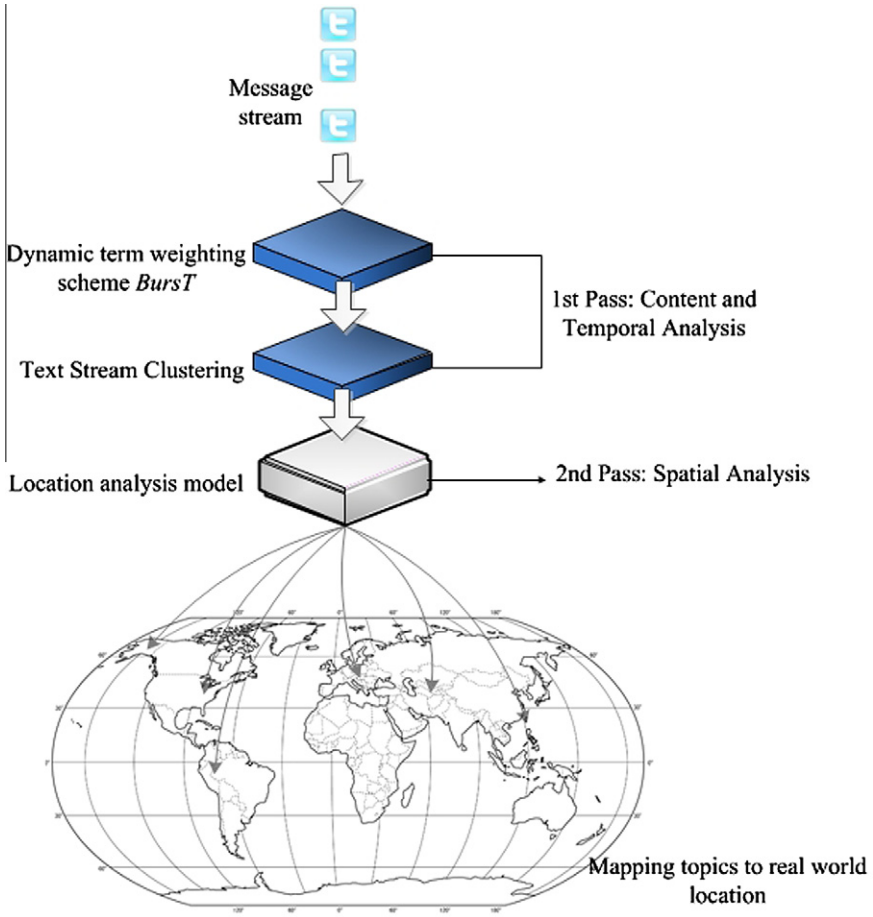


Fig. 1. Framework of system architecture.

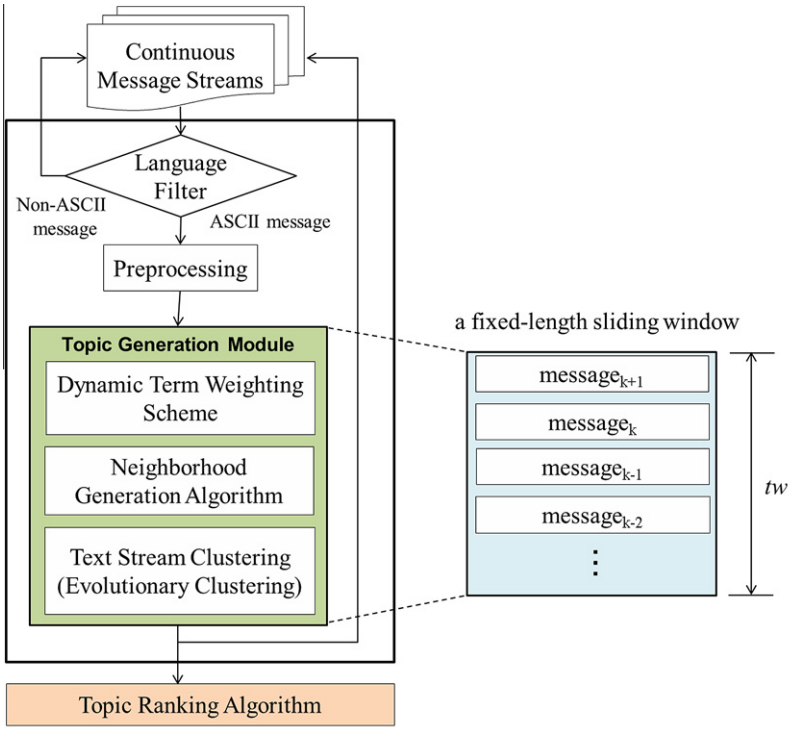


Fig. 2. Overview of proposed microblogging topic detection system.

lapsing. A typical approach for dealing with the problem is based on the use of so-called *sliding windows* (Bifet, 2010). As a result, the sliding window method is adopted for tackling the issue in this work, as shown in Fig. 3. Briefly speaking, the steps of sliding window technique include: (i) the insertion operation in which a new index entry is built when a message comes in, (ii) the message is reserved until its lifetime exceeds the fixed length of time window *tw*, and (iii) the deletion operation in which the message will be removed from memory.

2.3. Spatial analysis (2nd pass)

When event topics are detected, the next step is to analyze spatial distribution of them. In this work a location estimation method is utilized for estimating where the event occurs. The main idea behind our event detection approach is based on the characteristics of *spatial locality*. Spatial locality is described as a set of messages concerning some topic, which are highly densely located in a specific geographical area. Such a phenomenon is considered in our work to derive algorithms for estimating location of an event by using location feature vectors. A location feature vector records the location distribution of the topics at a specific time point. For example, the Christchurch earthquake in New Zealand took place on Feb 22, 2011. Once the news media broadcast the news, people in other places started to discuss about this event. The distribution of the users discussed the event was being expanded to other continents. Therefore the location feature vector is also changed over time as shown in Fig. 4. The algorithm for a spatial analysis model is shown in Table 1.

In Fig. 4, $occur_{n,m}$ denotes the number of occurrence of the mth location-feature in the nth topic and, t denotes the time of event occurrence. The location feature vector can be used to distinguish whether the event is a local topic or global topic. First we formulate the probability of a topic $topic_i$ which belongs to a specific location loc_j as Eq. (1).

$$p(loc_j|topic_i) = \frac{|occur_{i,j}|}{N_i} * \frac{1}{|loc_j|} \quad (1)$$

where the probability of the $topic_i$ belonging to a location loc_j is obtained by dividing the number of messages which contain loc_j in $topic_i$ ($occur_{i,j}$) by the total number of messages (N_i). The $1/|loc_j|$

Table 1
Spatial analysis model.

```
Algorithm: spatial analysis model
1:  INPUT: Array[0,I,...,m] [0,1,...,n]LFV //location feature vector
2:  OUTPUT: Array[0...n] locOfTopic
3:  Begin
4:    foreach topic[i]
5:      foreach locatian[j]
6:        totalNumOfOccur[i] ← totalNumOfOccur[i] + LFV[i][j]
7:      end for
8:    end for
9:    foreach tapic[i]
10:     foreach location[j]
11:       double locPropotion[i][j] ← LFV[i][j]/totalNumOfOccur[i];
12:       if(locPropotion[il[j] > θ) locOfTopic[j] ← "locaTopic : " +
          String(loc_j);
13:       then else continue;
14:   end for
15: locOfTopic[i]← "gtobalTopic";
16: end for
17: Return locOfTopic;
```

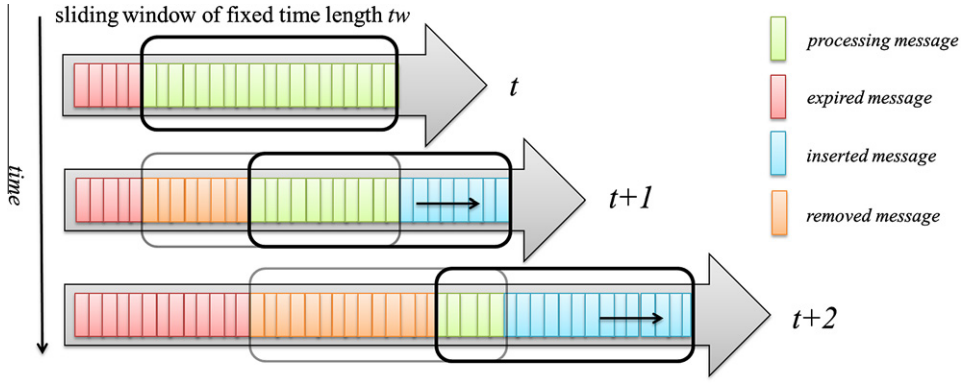


Fig. 3. A continuous message stream using sliding window model.

Time Zone	$Topic_1(t)$	$Topic_2(t)$	$Topic_3(t)$	$Topic_4(t)$		$Topic_{n-1}(t)$	$Topic_n(t)$
Athens	$occur_{1,1}(t)$	$occur_{2,1}(t)$	$occur_{3,1}(t)$	$occur_{4,1}(t)$		$Occur_{n-1,1}(t)$	$occur_{n,1}(t)$
Bangkok	$occur_{1,2}(t)$	$occur_{2,2}(t)$	$occur_{3,2}(t)$	$occur_{4,2}(t)$	• • •	$occur_{n-1,2}(t)$	$occur_{n,2}(t)$
Beijing	$occur_{1,3}(t)$	$occur_{2,3}(t)$	$occur_{3,3}(t)$	$occur_{4,3}(t)$		$occur_{n-1,3}(t)$	$occur_{n,3}(t)$
Hawaii	$occur_{1,4}(t)$	$occur_{2,4}(t)$	$occur_{3,4}(t)$	$occur_{4,4}(t)$		$occur_{n-1,4}(t)$	$occur_{n,4}(t)$
	⋮					⋮	
Central Time (US & Canada)	$occur_{1,m-1}(t)$	$occur_{2,m-1}(t)$	$occur_{3,m-1}(t)$	$occur_{4,m-1}(t)$	• • •	$occur_{n-1,m-1}(t)$	$occur_{n,m-1}(t)$
Hong Kong	$occur_{1,m}(t)$	$occur_{2,m}(t)$	$occur_{3,m}(t)$	$occur_{4,m}(t)$	• • •	$occur_{n-1,m}(t)$	$occur_{n,m}(t)$

Fig. 4. Location feature vector.

is the penalty factor to penalize a topic which is widely discussed in many places. In addition, a candidate location (as shown in Eq. (2)) can be determined by the maximum of probability for $topic_i$.

$$candiLoc(topic_i) = \arg\max_{loc_j}\{p(loc_j|topic_i)\} \quad (2)$$

In Eq. (2), a candidate location of the topic is calculated. Subsequently the rule for determining whether the topic is a local topic or global topic can be formulated as Eq. (3).

$$Loc_i = \begin{cases} candiLoc(topic_i), & \text{if } p(candiLoc(topic_i)|topic_i) > \theta \\ \text{“}globalTopic\text{”}, & \text{otherwise.} \end{cases} \quad (3)$$

In Eq. (3), a topic would be regarded as a local topic if the probability of candidate location exceeds the threshold θ. The cut-off point θ represents a tradeoff between the level of sparsity and concentricity for a given topic. For setting cut-off point θ, the higher the threshold value is chosen, the more concentricity of the topic is required. Hence, an event is considered as a local event if the distribution of messages is sparse, otherwise it is a global event.

Finally, the geospatial distribution of the topic can be mapped to a real world map for visualization. The spatial analysis model is designed to support identifying where the event happened. Even once a local event has been become to a global event, we can still trace back to its early state.

3. The proposed density-based online clustering method

In this section, we describe the technical details of the proposed density-based online clustering method. Our work starts with investigating the solution for detecting topics and tracking events about the interests, hot news topics, and preferences of people from text information sources of microblogging services. In this work, an algorithm using a density-based method is developed for mining microblogging message streams. The purpose of our approach is to effectively detecting and grouping emerging topics from the user-generated content in a real-time or specified time slot. On the other hand, for tackling a key challenging issue in mining the microblogging messages, we attempt to analyze the real-time distributed messages and extract significant features of them in a dynamic environment. We propose a novel term weighting method, called *BursT*, using a sliding window technique for weighting message streams. This method was proven to be capable of dealing with concept drift problem, being able to detect context changes without being explicitly informed about them.

3.1. The density-based clustering approach

As the temporally-ordered messages streaming into the system, the next step is to incrementally gather messages into thematically topics. For such an information gathering process, one of the main difficulties is figuring out the meaning and value of those fleeting bits of information for mining the text streams. The challenge goes beyond filtering out spam, though that's an important part of it. Microblogging messages may lose their value within minutes of being written. Therefore, the system should be able to quickly group them into clusters which are evolving over time. Meanwhile, the continuous evolution of clusters makes it essential to be able to swiftly identify new clusters in the data. That is, the algorithm has to deal with lots of external dynamic changes, i.e. various updates occur and topic shift (i.e. concept drift) issues, etc. In order to achieve this goal, we have to provide an effective solution in which online clustering operation can be well performed in mining the microblogging text streams.

3.1.1. The considerations for utilizing density-based clustering approach

The reasons of adopting a density-based clustering approach in this work are described as follows:

(1) Messages collected from microblogs normally contain lots of noises. Once mining microblogging messages, the clustering algorithm should perform its best to filter out noises in processing the contents. Density based clustering groups data based on their density connectivity and treats noises as outliers which would not be involved in any cluster
(2) Density-based clustering techniques are capable of detecting arbitrary-shaped clusters.
(3) There is no assumption about the number of clusters with fixed or flexile parameter of k (i.e. topic), and it is thus unsuitable for some real world applications in the problem domain, especially in dealing with the topic detection task with dynamic topics around the world.

Due to the dynamic natures mentioned above, it is highly desirable to perform data updates incrementally. Thus, in this work a density-based clustering based on the algorithm of IncrementalDBSCAN (Ester, Kriegel, Sander, Wimmer, & Xu, 1998) was used for our system development. IncrementalDBSCAN is an efficient algorithm which is based on DBSCAN for mining data with density-based connectively (Nguyen-Hoang, Hoang, Bui-Thi, & Nguyen, 2009; Sun and Hu, 2011; Wen, Nie, & Zhang, 2002). The technique and operational application of IncrementalDBSCAN are described in details in the following subsections.

3.2. The IncrementalDBSCAN algorithm

In this work, we adopted a density-based clustering approach called IncrementalDBSCAN to against noises, instead of using spam classifier to determine uninformative message subjectively. The key idea of density-based clustering is that for each object of a cluster, the neighborhood of a given radius (*Eps*) has to contain at least a minimum number of objects (*MinPts*) to form as a basic unit of density region. The dynamic operations for insertion can be divided into *Noise*, *Creation*, *Absorption* and *Merge* conditions and deletion has three different cases of *Removal*, *Reduction* and *Potential Split*. Each case is judged from the properties of connectivity, and enables us to maintain the status of the cluster dynamically.

According to the theory of IncrementalDBSCAN clustering method, the shape of clusters will change over time when a message being inserted or a victim message being deleted from sliding window with its message density properties. Certainly the less density area would not be a topic, because of the distances between messages are long according to the calculations of temporal text similarity. Meanwhile, text stream cluster algorithm will generate several clusters at each time, due to its natural dynamics.

Essentially, the key idea of density-based clustering is that for each object of a cluster the neighborhood of a given radius (*Eps*) has to contain at least a minimum number of objects (*MinPts*), i.e. the cardinality of the neighborhood has to exceed some threshold. The core objects will be established as they reach the statement, and then only the part of *affected objects* (see Fig. 5) will be updated with the incremental version. The core objects in *NEps(p)* can be defined as $UpdSeed_{\mathrm{Ins}}$ and $UpdSeed_{\mathrm{Del}}$ for operations of insert and delete, where *NEps(p)* is the set of *Eps*-neighborhood of *p*.

3.3. Temporal text similarity

When a message arrived, the similarity should be calculated with its neighbors. In our approach, we proposed a formula called

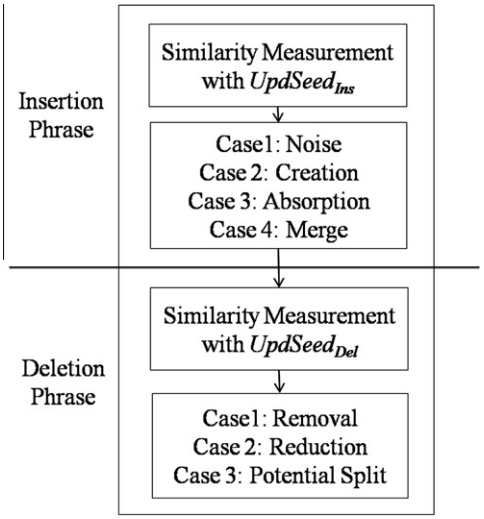


Fig. 5. Two phrases of text stream clustering in IncrementalDBSCAN approach.

temporal text similarity which combines content and temporal dimensions. Supposed that we have an incoming message m_a, and a neighbor massage m_b, the similarity between m_a and m_b can be denoted as follows:

$$Sim(m_a, m_b) = \cos(m_a, m_b) * tp(m_a, m_b) \tag{4}$$

$$\cos(ma, mb) = \sum_i \frac{m_{ai} * m_{bi}}{|m_a||m_b|} \tag{5}$$

$$tp(m_a, m_b) = e^{\frac{\zeta|t_a - t_b|}{W}} \tag{6}$$

In similarity function $Sim(m_a, m_b)$, the cosin similarity is used for content-based similarity measurement. In order to make a consideration of temporal information, the **temporal penalty** $tp(m_a, m_b)$ is an exponential distribution which can reduce the similarity if two document with a long time distance, the temporal penalty is high when the distance $|ta - tb|$ is close and vice versa. The ζ is a parameter which can adjust the temporal decay rate.

4. Design of a dynamic term weighting scheme for adapting to changes in messages

As mentioned previously, in this work our solution to tackle the issues is based on the utilization of a *sliding window*. As the context is known to vary in time, the learner trusts only the latest examples – this set is referred to as the *window*. Data samples are added to the window as they arrive, the oldest samples are deleted from it. In our solution, the window is being of fixed size, and the oldest sample will be dropped whenever a new one comes in. Meanwhile, once changes of a concept have been detected, the system should be able to discard out-of-date examples and clusters (e.g. time window). In this work, we utilize a dynamic weighting scheme to discriminate the event messages, and cope with the concept-drift issues (Khalilian & Mustapha, 2010; Widmer & Kubat, 1996) in a dynamic environment. Once the concepts behind the messages evolve with time, the underlying cluster structures in our system will also significantly change with time. Under such a circumstance, the system should be able to be adapted itself to supporting topic and concept drifting in the microblogging text streams. We developed a novel term weighting method, called *BursT*, using sliding window techniques for weighting message streams (Lee et al., 2011). The experimental results show that our weighting technique has an outstanding performance to reflect the occurrence of concept drifts in tweets.

4.1. The dynamic term weighting scheme

For microblogs, the corpus tends to be dynamic as new items always being added and old items being changed or deleted. Therefore, an ideal term weighting scheme for mining microblogging messages should subtly reflect the changes over time and quickly assign proper weights in such a dynamic environment. In this section, we firstly describe the sliding window model, which is associated with the development of our weighting method, and then explain the weighting mechanism.

4.1.1. The design strategy for term weighting

To process texts with a chronological order, a fundamental problem we concerned is how to find the significant features in text streams. Specifically, the trends of concept are often not stable but change with time, which is also known as *concept drift*. Under such a circumstance, the design of weighting scheme for microblogging message should be constantly updated. Here we apply the term weighting scheme *BursT* which was proposed in our previous work (Lee et al., 2011). The experimental results indicate that has a better performance in weighting words of microblogging messages than *incremental TFIDF* (Brants, Chen, & Farahat, 2003) and *TFPDF* (Bun & Ishizuka, 2002) techniques.

The word *burst* is defined as an unusual number of frequently posted messages happened in a short time. Fig. 6 shows a three-phase of word categories occurred in microblogging messages. In Fig. 6, *uninformative word* means that a word rarely occurred in the sliding window, such as an oral word. The axis df shown in Fig. 6 represents the value of document frequency. If the word was occurring very frequently but with lower burstiness, they could be recognized as *common word* or *social word*, such as "haha" and "lol". If a word has a higher burst than expectation within a certain range of document frequency, we will highlight its importance for weight design in the sliding window.

Accordingly, our strategy in determining *BursT* value is that a heavier weight is achieved by a higher burstiness, in which some word occurs frequently in the window. Thus, the *BursT* weighting formula is shown in Eq. (7):

$$weight_{w,t} = BS_{w,t} * TOP_{w,t} \tag{7}$$

where the weight of the word w at time t will be constituted by two factors: *BS* (*Burst Score*) and *TOP* (*Term* Occurrence Probability). For calculating *BursT* weights of single words, each word w is recorded as a quartet $\langle w, at_{w,t-1}, n_{w,t}, E(ar_{w,t})\rangle$, $at_{w,t-1}$ represents the last time word w arrived, $n_{w,t}$ counts the total number of word w appeared in our system, and $E(ar_{w,t})$ is a long time cumulative expectation of arrival rate to the word w. The detailed description of the weighting factors will be discussed in the following subsections.

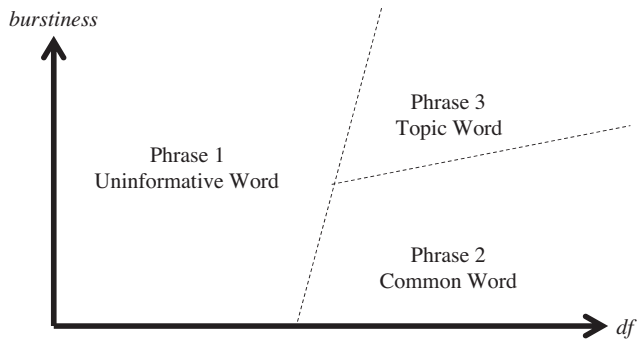


Fig. 6. Word categories in microblogging texts.

4.1.2. The BS weighting factor

The interval of arrival time between messages can be transformed into arrival rate for many streaming data applications. If the feature of some message arrives with short intervals incessantly, the feature representing the importance of a message may be more useful. Suppose there is a feature word w occurs in message sequence $\{m_{w,1}, m_{w,2}, m_{w,3} \ldots m_{w,t}\}$, and each message has a specified arrival time $at_{w,t}$. We can then define the arrival rate $ar_{w,t}$ for current message $m_{w,t}$ by the formula shown in Eq. (8):

$$ar_{w,t} = \frac{1}{at_{w,t} - at_{w,t-1} + 1} \tag{8}$$

In Eq. (8), if $t = 1$, the interval value becomes zero because w is a brand new word in the system. The arrival rate $ar_{w,t}$ represents the reciprocal type of arrival gap $(at_{w,t} - at_{w,t-1})$ which could be normalized between 0 to 1. In order to reflect long time expectation of arrival rate of the word, the mean value for each word is calculated in an incremental manner.

$$\mu_n = \mu_{n-1} + \left(\frac{1}{n}\right)(x_n - \mu_{n-1}) \tag{9}$$

$$E(ar_{w,t}) = \mu_{w,t} = \mu_{w,t-1} + \left(\frac{1}{n_{w,t}}\right)(ar_{w,t} - \mu_{w,t-1}) \tag{10}$$

In this work we apply incremental mean (Finch, 2009) (i.e. Eq. (9)) in our weighting scheme to formulate equations of insertion (i.e. Eq. (10)), where $ar_{w,t}$ is the new arrival rate of the word.

$$RMOar_{w,t} = \sum_{n=1}^{k} \frac{ar_{w,t-n}}{k} \tag{11}$$

Then the burst score is calculated as below:

$$BS_{w,t} = \max\left\{\frac{RMOar_{w,t} - E(ar_{w,t})}{E(ar_{w,t})}, 0\right\} \tag{12}$$

Therefore, we regard $ar_{w,t}$ as the current observation result, to compare with expected value $E(ar_{w,t})$ of the word w at tth arrival. In addition, we derive a formula Eq. (11) in which residual is the deviation between observation and expectation values. It should be noted that the result of Eq. (12) would not always be positive if the observation result is less than expectation value. In such a case, we define the word as a "falling word" at that time, and enable *BS* factor to be zero.

4.1.3. The TOP weighting factor

The second consideration in *BursT* weighting scheme is *TOP* (*term* occurrence probability) factor, which is formulized by the proportion of the term in the sliding window. For the operation of mining hot news topics from messages, if a word occurs in more messages, it is more likely to be a trending topic. Thus, the term occurrence probability corresponding to the word w at tth arrival is formulated as below:

$$TOP_{w,t} = P(w_t|c_t) = \frac{|\{m : w_t \in c_t\}|}{|c_t|} \tag{13}$$

where *TOP* represents the probability of the word occurrence in the sliding window, and c_t denotes the message collection in the corpus collected from the time *t-tw* to current time. This factor would enable the weight of the word to grow with its occurrence frequency in messages, for identification of trending topics.

5. Experimental results

We designed two sets of experiments to evaluate (i) the effectiveness of *BursT* weighting method, and (ii) the validity of our event detection system, by means of analyzing the spatio-temporal impacts of some selected events detected by our system as our case studies. The two sets of experiments are described in the following subsections.

5.1. Experiments with BursT weighting method for event detection

In order to examine the system performance in reflecting the concept drift of words, we selected "Chile's Rescued Miners" event as our case study. Our experiment started with the first miner was rescued at 2010/10/13 11:11 (GMT +8:00), until all miners were rescued at 2010/10/14 20:56 (GMT +8:00). Fig. 7 indicates that the intensity of inter-arrival gap the feature word "chile", and it suddenly dropped at Oct 13 (GMT +8:00) when the event was happening.

Subsequently, we compared the weighting values of *BursT* and *TFIDF* methods, as shown in Fig. 8. We found that the incremental *TFIDF* can't reflect the actual trends in sliding window algorithm, but *TFPDF* and our approach performed well in topic words. However, in the outcome of oral word analysis, we demonstrate the word "lol", which both has a high density of collection and arrival rate, as an example, and obviously it might not be suitable to define "lol" as a valid feature. It is worth mentioning that some popular oral words might be easily over weighted in *TFPDF* because it places too much emphasis on document frequency. As shown in Fig. 8(d) and (e), the weighted number of "lol" in *TFPDF* is still higher than in incremental *TFIDF* even the event is still on the fly.

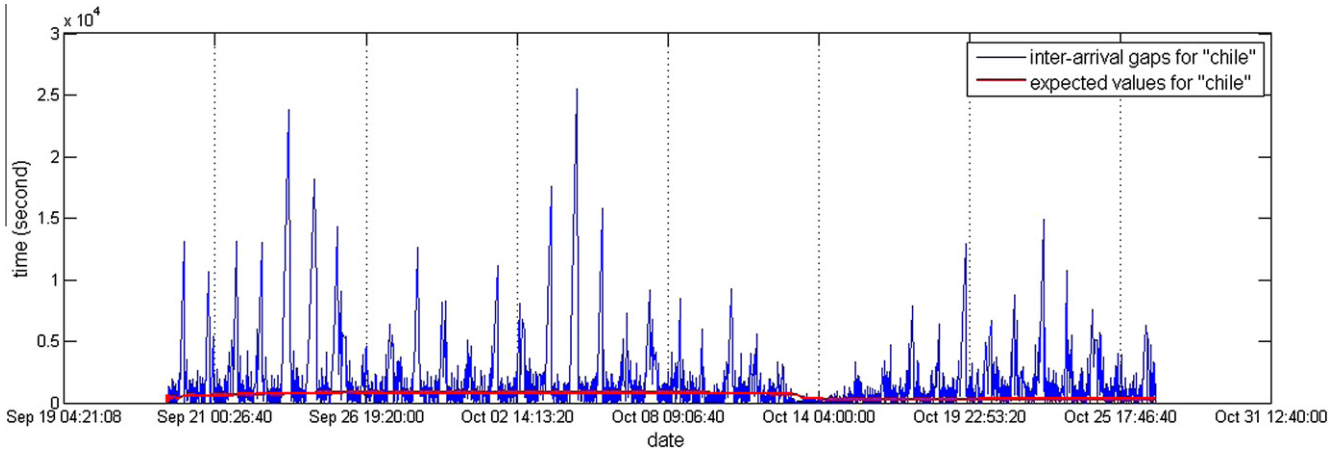


Fig. 7. Inter-arrival gaps using the feature word "chile".

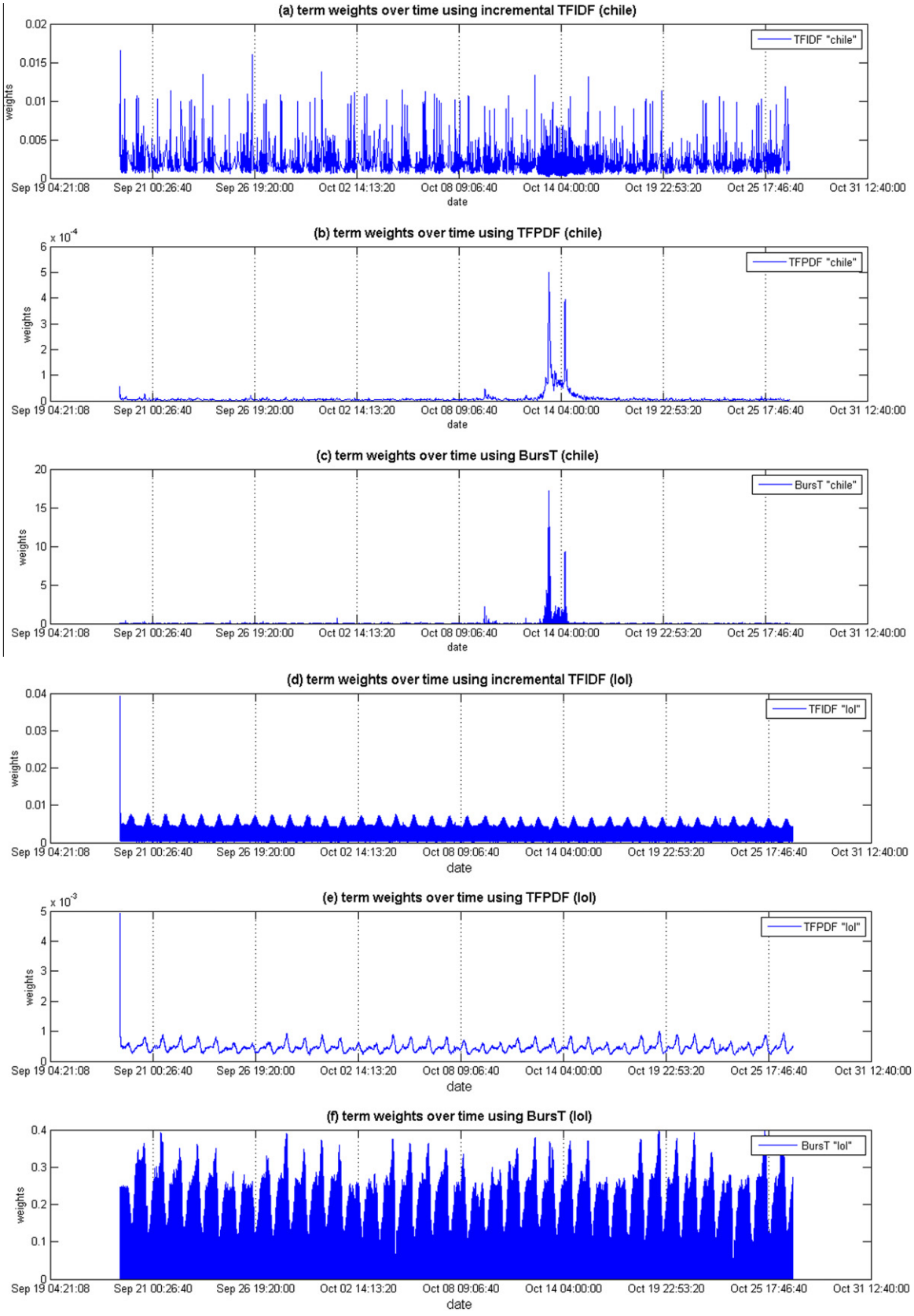


Fig. 8. Evaluation of *BursT*, incremental *TfIDF* and *TFPDF* weighting techniques.

5.2. Temporal and spatial analysis for real-world events

This set of experiments aims to find out the feasibility of our model used in situational awareness by examining experimental spatio-temporal results in several practical case studies. To identify the validity of our framework, we have utilized our developed approach to perform experiments on the real-world events. In terms of temporal dimension, we performed experiments on our real-time event detection model, with the following parameter settings: *Eps* = 0.4, *MinPts* = 15, and the length of sliding window *tw* is set to two hours. Once events are detected, each message in the topic will be analyzed for acquiring the most possible place the event originally happened. For our experiments, geographical nouns and proper nouns are collected to construct a lexicon, including geospatial named entities such as Christchurch, Fukushima, etc, which are used to identify the names of locations associated with specific events in the Twitter messages. After that, these messages would be mapped to the Google Map, providing the resulting information about geospatial distribution of the event topic in the world.

5.2.1. Data source

In this set of experiments, a large number of microblogging posts were collected from Twitter during a period of time. These samples were collected through Twitter Stream API (*statuses/sample* method) to get random samples from all public messages. After filtering out non-ASCII tweets, some available tweets had been utilized as our data source. Further we partitioned messages into unigrams and removed the substring "RT @username:". Then, the stopword list includes thirteen different languages such as English, German and French since the collected tweets contain multilingual texts. Also, all capital letters in each tweet were converted into lowercase.

For spatial analysis, perhaps the utilization of latitude and longitude data is the most precise way for locating messages. However, according to our preliminary statistics on 270,852 sample tweets, we found that approximately 66,565(24%) tweets have no time-zone information from their user profile, and 268,831(99%) tweets have no latitude and longitude information through Twitter Stream API. As a result, the way of latitude and longitude informa-

Table 2
Sample detected events by our system for spatio-temporal analysis.

Event type & identifier	Event	Ground truth	Estimated location (time-zone based method)	Estimated location (content based method)
Others no.#89	Congresswoman Shot in Arizona (January 09)	Arizona	Eastern Time (US & Canada)	Arizona
Others no.#518	TV program "The Voice of Holland" (January 22)	Holland	Amsterdam	Holland
Politics no.#651	State of the Union (SOTU) Address (January 26)	Washington	Eastern Time (US & Canada)	Washington
Politics no.#1411	Egyptian President Stepped Down (February 12)	Egypt	Global Event	Egypt
Entertainment no.#1529	The Grammy Awards (February 13)	Los Angeles, California	Pacific Time (US & Canada)	Los Angeles
Disaster no.#2607	Christchurch Earthquake (February 22)	New Zealand (Christchurch)	Auckland	Christchurch
Politics no.#2552	Libya air strikes (February 22)	Libya	Global Event	Libya
Entertainment no.#3511	The Oscars Awards (February 28)	Los Angeles, California	Global Event	Hollywood
Disaster no.#4188	Magnitude 9.0 Japan Earthquake (March 11)	Japan (Miyagi)	Tokyo	Japan, Tokyo
Others no.#4487	Nate Dogg R.I.P. (March 16)	California	Central Time (US & Canada)	California
Others no.#4891	Elizabeth Taylor R.I.P. (March 23)	Los Angeles, California	Eastern Time (US & Canada)	California
Disaster no.#5584	Magnitude 7.4 Japan Aftershock (April 07)	Japan (Honshu)	Global Event	Japan, Honshu
Others no.#6486	Royal wedding (April 29)	England (London)	Pacific Time (US & Canada)	England London
Politics no.#6645	Osama Bin Laden killed by U.S. (May 2)	Pakistan (Islamabad)	Eastern Time (US & Canada)	America, Pakistan
Entertainment no.#7202	Eurovision Song Contest 2011 (May 15)	Germany (Düsseldorf)	London	Azerbaijan

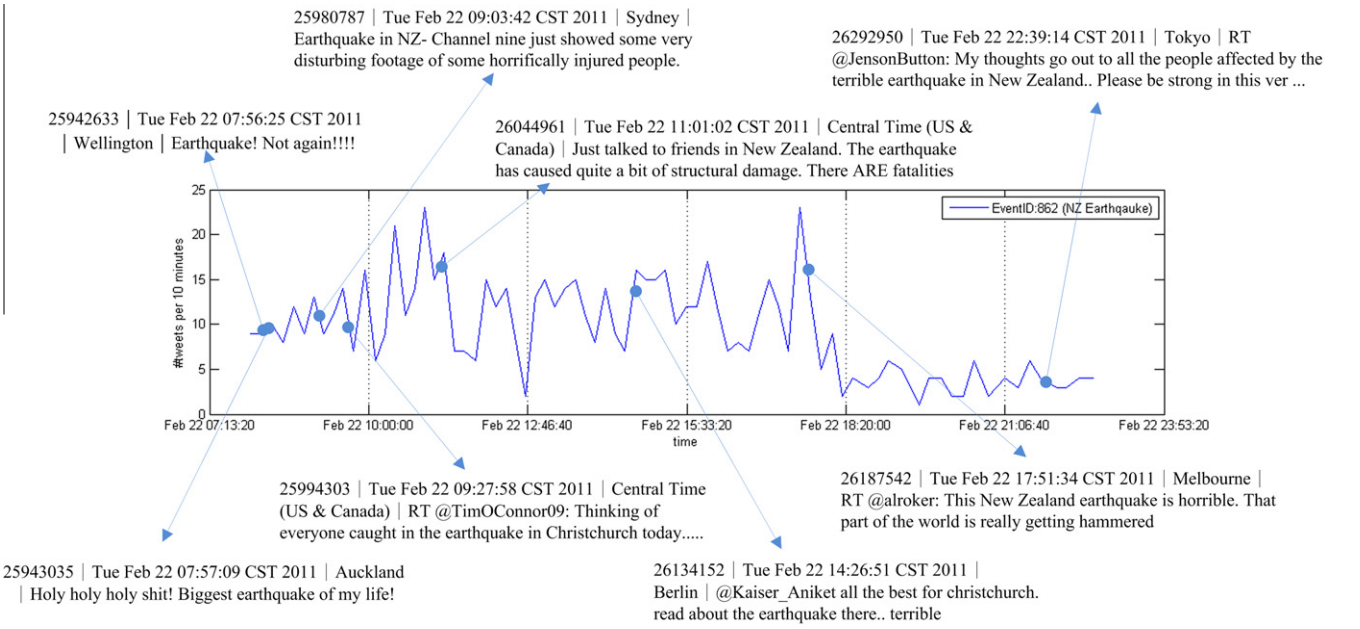


Fig. 9. Experimental result: temporal analysis of "Christchurch earthquake on Feb 2011" event.

tion in tweets is not well suited for detecting popular geographical events in our work.

5.2.2. Analysis of temporal locality

In our topic detection system, we assumed that each topic has characteristics of temporal locality. It means that a topic would be discussed by tweets during a period of time. The reason we use the ways for mining topics rather than using keyword-based text retrieval methods is due to that such techniques can group relevant posts based on similarity of messages, avoiding missing valuable messages.

5.2.3. Experiments and results

In the experiments, we actually examine our system function by analyzing the real-world events detected by the system, in terms of temporal and spatial impacts of the events. Several sample events detected by our system for spatio-temporal analysis are selected and organized, as shown in Table 2. Among these events, we particularly select three events to conduct experiments to perform the temporal and spatial analysis. These experiments are described below.

5.2.3.1. Experimental event I: Christchurch earthquake on Feb 2011. In this work, the event of Christchurch earthquake on Feb 2011 was taken for our first case study. The event happened in Christchurch (+GMT 12:00), and we collected posts in Taiwan (+GMT 8:00). The results of spatio-temporal analysis are being described as follows. The temporal analysis of "Christchurch earthquake on Feb 2011" on Twitter is illustrated in Fig. 9. The spatial analysis of the event for location estimation by extracting time zone data and geospatial keywords from contents in tweets is illustrated in Fig. 12(a) and (b). Fig. 12(c) illustrates the event evolution based upon multiple dimensions (i.e. time, geospatial keyword, and number of messages), which is used to explore the interplay among the three dimensions by gathering tweets sampling per ten minutes in extracting insightful summaries of observations.

- Discussion (experimental event I)
  (1) In Fig. 9, the message number of 25942633 and 25943035 are obviously the original participants of the earthquake related tweets. They posted messages about local events like a journalist so we can acquire the first hand real-time information. Although there are no geolocation words in their contents, these posts would be correctly aggregated

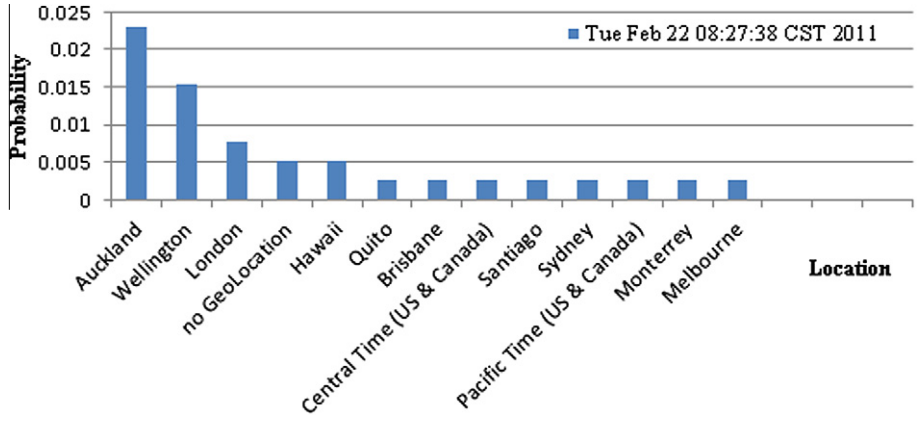


Fig. 10. Topic location probability distribution of "Christchurch earthquake"on Feb 22 08:27:38 2011.

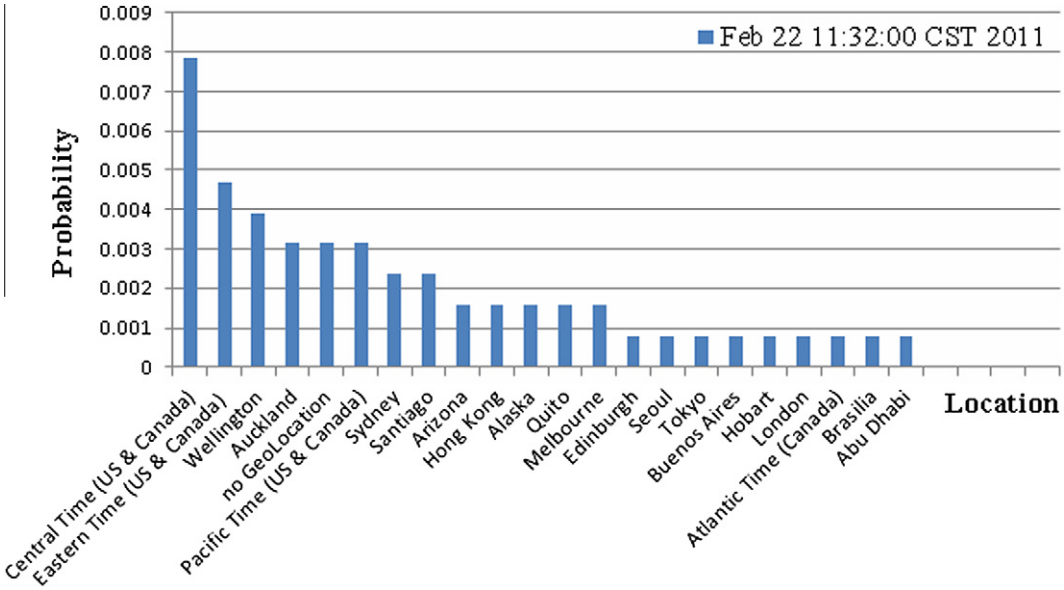


Fig. 11. Topic location probability distribution of "Christchurch earthquake"in Feb 22 11:32:00 2011.

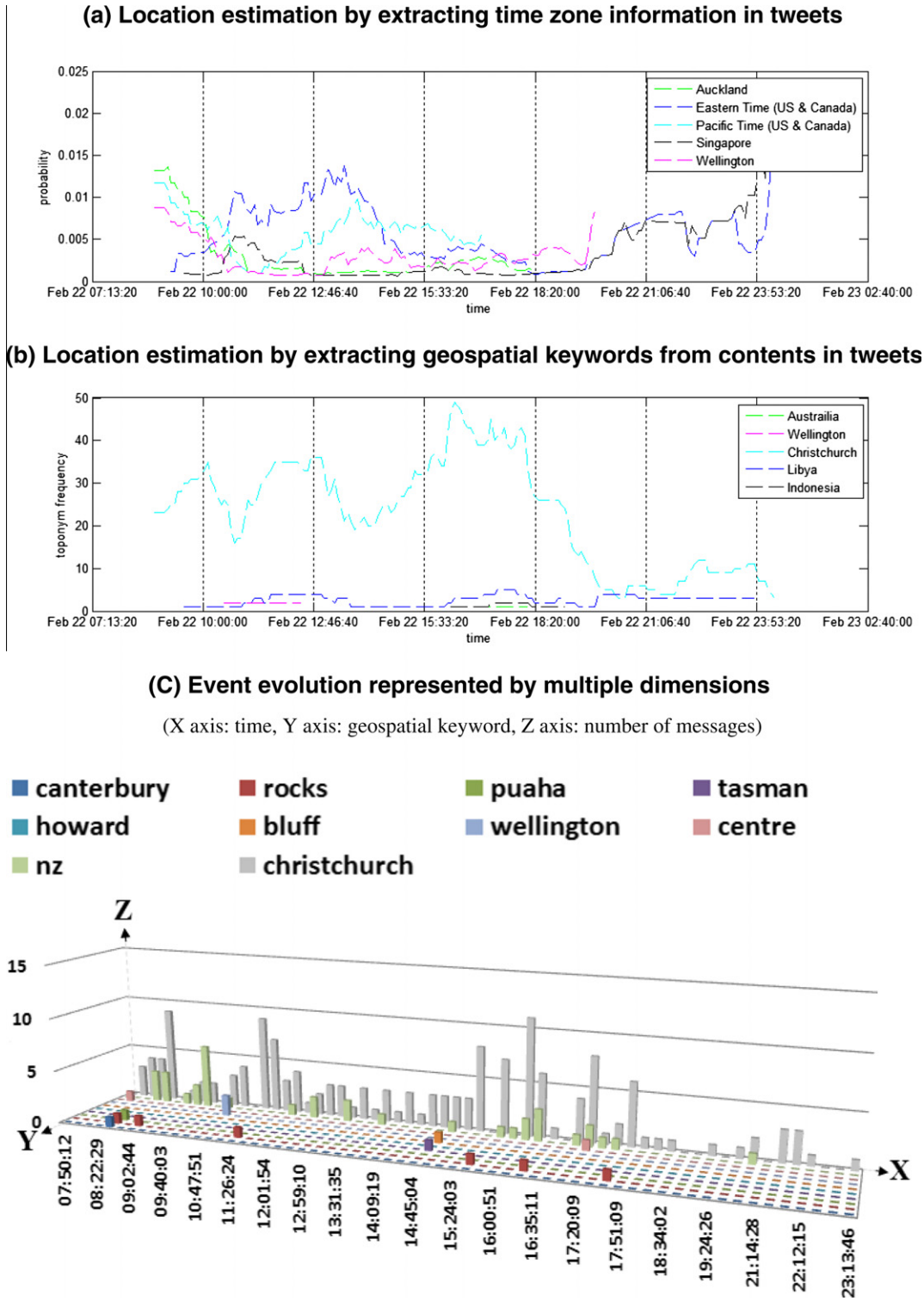


Fig. 12. (a) Event I: location estimation by extracting time zone information in tweets. (b) Event I: location estimation by extracting geospatial keywords from contents in tweets. (c) Event evolution representation based upon three dimensions.

in the same event through a simple word "earthquake". This is due to the term weight of "earthquake" was computed as 1.145 by our algorithm, enabling the term to be recognized as a bursty word at that time. Such type of bursty words can benefit catching concept drift of a topic. In our case study, only a word "earthquake" was identified as a bursty word when the earthquake occurs. After a while, if there existed a post containing both "earthquake" and "Christchurch" bursty words, which may establish neighborhood relations with the messages only contain the word "earthquake" or "Christchurch". Therefore we acquired a concept drift associated with the event from message streams.

(2) In this work the detected event topics will be identified to understand their geospatial impacts by utilizing the time zone data or extracting geospatial keywords in the content of each post. In the early stage of the event, we found that

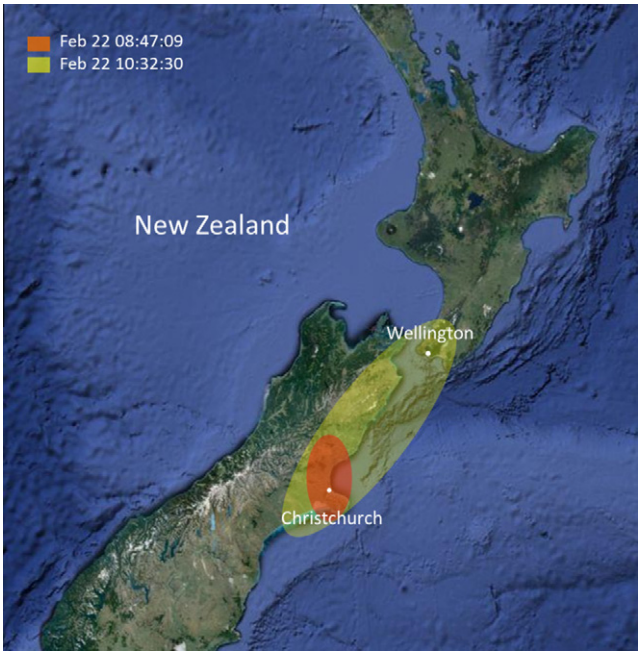


Fig. 13. Event visualization (event 1).

most of related microblogging messages appeared to be located central to Auckland and Wellington cities which are metropolitans near Christchurch in New Zealand (Fig. 10). Such a phenomenon is regarded as the characteristics of spatial locality among tweets, which is appearing at the early stage of an event. As time passes by, the news was broadcasted, and the related tweets started to spread to other cities as shown in Fig. 11. For a while, the country and cities containing the largest number of related discussion are changed to the ones in the United States. This is because most Twitter users were located in America then. The opinions in these messages were discussing about New Zealand earthquake, and some tweets were still popular but not first-hand information. Hence, we can utilize the characteristics of spatial locality, tracing back to the earlier stage of the event to find out which cities may be the most possible area where the event originally took place.

(3) As we can see in Fig. 11, once the USA is becoming the most massive region for the discussion, the probability will be penalized because the event has been widely discussed in many places around the world. In order to get insight into the event, we mapped messages related to the event into the Google Map for visualization, as illustrated in Fig. 13.

5.2.3.2. Experimental event II: Japan earthquake on March 2011. Subsequently, the event of Japan earthquake on March 11, 2011 was taken for our case study. The results of spatio-temporal analysis are being described below. For this experiment, we first selected, for each day, only the cities that appear in the Twitter messages. The results of detecting the "Japan earthquake on Mar 11, 2011" have been investigated. This event happened at 13:46 (GMT +8:00), and the first related post we got is the message of "Quake!" from Tokyo at 13:50:53 (GMT +8:00). The topic cluster was established by our system at 14:03:28 (GMT +8:00). The temporal analysis of "Japan Eearth quake on March 2011" on Twitter is illustrated in Fig. 14. The spatial analysis of the event for location estimation by extracting time zone data and geospatial keywords from contents in tweets is illustrated in Fig. 15(a) and (b). Fig. 15(c) illustrates the event evolution representation based upon the dimensions of time, geospatial keyword, and the logarithm of the number of messages.

- Discussion (Experimental event II)

(1) In this case study, we marked the three time points in the event "Japan earthquake". At the first time point the event related to earthquake was detected by our system. After that, at the second time point, news about the tsunami caused by Japan earthquake may head toward Hawaii start to broadcast. At the third time point, the Japan government confirmed the radiation leak at Fukushima nuclear plants. In addition, following the event, the information diversity was gently increased. This demonstrates that our system have the capacity for detecting concept drift, since the system did not separate them into three different topics.

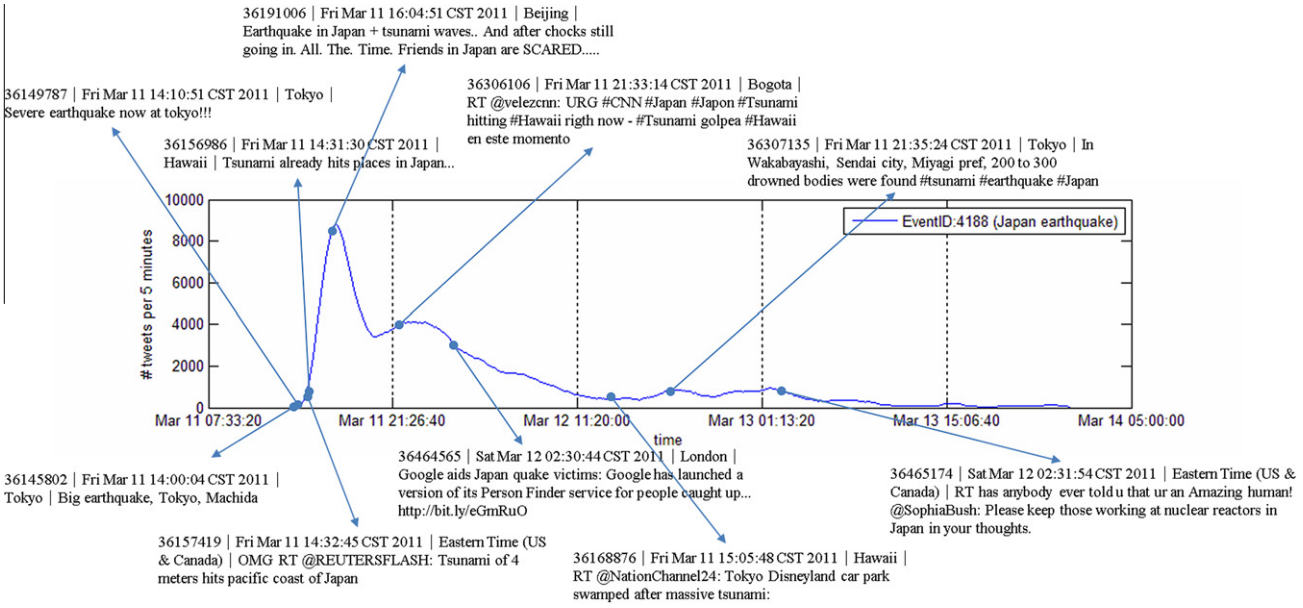


Fig. 14. Experimental result: temporal analysis of "Japan earthquake on March 2011" event.

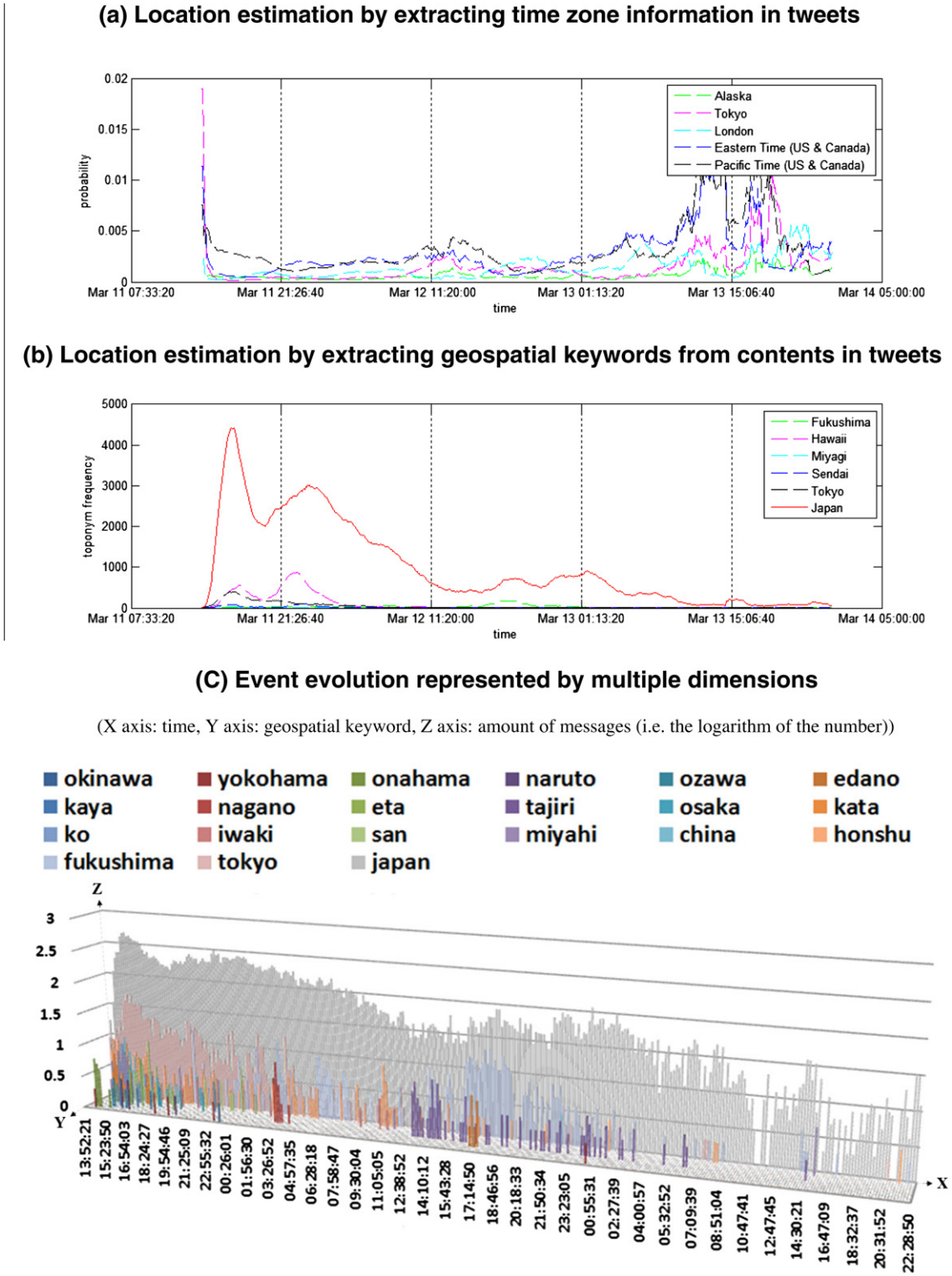
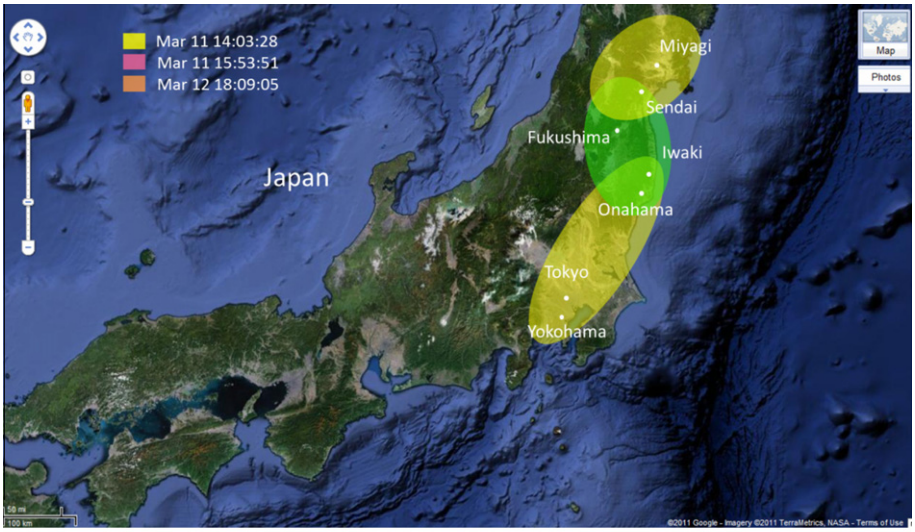


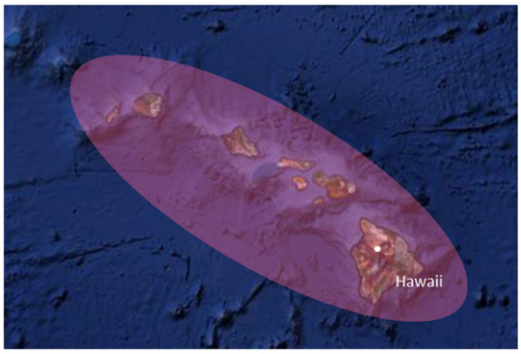
Fig. 15. (a) Event II: location estimation by extracting time zone information in tweets. (b) Event II: location estimation by extracting geospatial keywords from contents in tweets. (c) Event evolution representation based upon three dimensions.

(2) One of the most interesting features of IncrementalDBSCAN method is its "*transitive*" characteristics (Ester et al., 1998). The transitive characteristics allow users to easily incorporate relevant messages into related topic clusters. In this case, such an advantage with our method for event detection can be proved. Suppose that we have three sets of documents: A, B and C. If there exists a situation that A is density-reachable to B, and B is density-reachable to C. It implies that a transitive relation is established between A and C. In this event, for instance, once an earthquake event happened, the topic cluster may contain lots of discussion groups about "earthquake". After a while, some messages about "tsunami" appear, in other words, the earthquake event starts drifting its concept imperceptibly to a

"tsunami" event, and then move to the topic of "nuclear". Transitive characteristics can help create a hidden connection, which indicates that the messages of "earthquake" and "nuclear" can thus be grouped to the same discussion topic related to the original event. Such a transitive absorption capacity is useful to adapt the effects of concept drift, and prevent the discussion topics transfer to different event. For visualization, we mapped messages related to the event into the Google Map, as illustrated in Fig. 16.



(a) Event visualization(event 2(i))



(b) Event visualization(event 2(ii))

Fig. 16. (a) Event visualization (event 2(i)), and (b) event visualization (event 2(ii)).

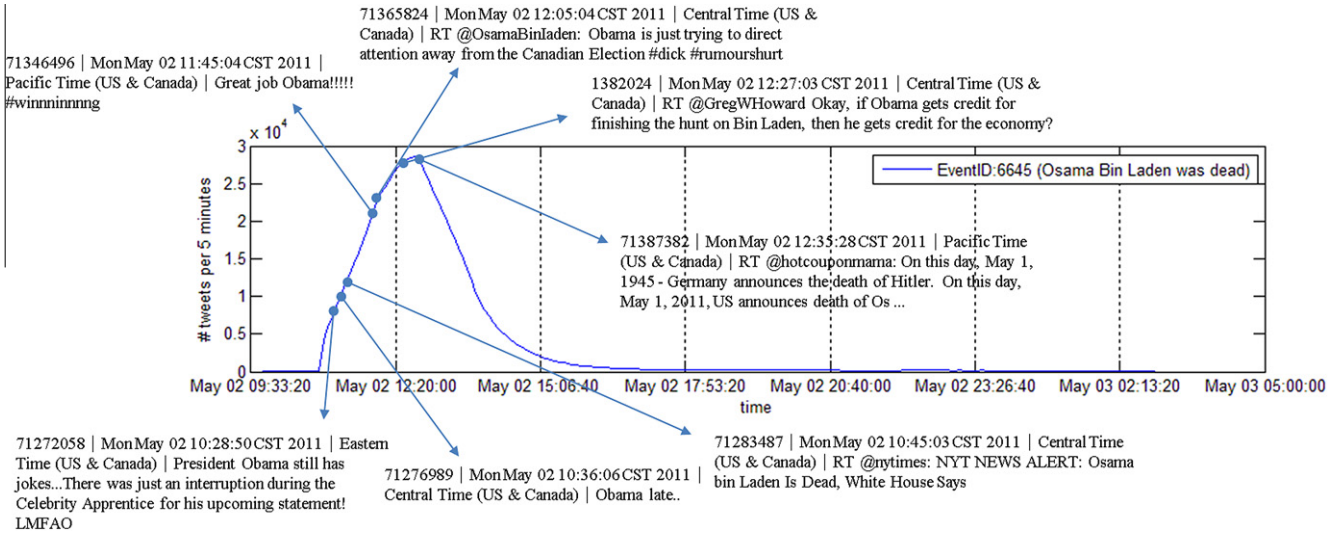


Fig. 17. Experimental result: temporal analysis of "President Obama Delivers Statement on Death of Osama Bin Laden" event.

5.2.3.3. Experimental event III: President Obama delivers statement on death of Osama Bin Laden. The final sample event discussed in this paper is regarding the event "President Obama Delivers Statement on Death of Osama Bin Laden" on May 2, 2011. The results of spatio-temporal analysis are being described below, and related discussions of the event on Twitter are shown in Fig. 17. We take this event as an example for observing how our approach can be used to monitor such crowd-sourced summaries to supplement situational awareness by means of our purely non-political tool. We believe that this characteristic of the data does not introduce any bias in our experiment since our techniques do not consider the tweets' political beliefs in the data collection process. For this experiment, we first selected, for each day, only the cities that appear in the Twitter messages. The results of detecting the event "President Obama Delivers Statement on Death of Osama Bin Laden" have been investigated. This event happened at 22:45 (GMT +8:00) May 01. The topic cluster was established by our system at 09:46:28(GMT +8:00) May 02. The resulting temporal analysis of "President Obama Delivers Statement on Death of Osama Bin Laden" event on Twitter is illustrated in Fig. 17. The spatial analysis of the event for location estimation by extracting time zone data and geospatial keywords from contents in tweets is illustrated in Fig. 18(a) and Fig. 18(b). Fig. 18(c) illustrates the event evolution representation based upon dimensions of time, geospatial keyword, and number of messages factors.

Prior to the presentation of President Obama in this event, the content of the presentation was kept as a top secret. As a result, in Twitter most of the related tweets attempted to discuss and guess the possible content of the report being announced then. Initially, the event was limited to the United States, and so most of the related tweets were located in the USA. As soon as President Obama started his talk at 10:30, the geographical areas containing the

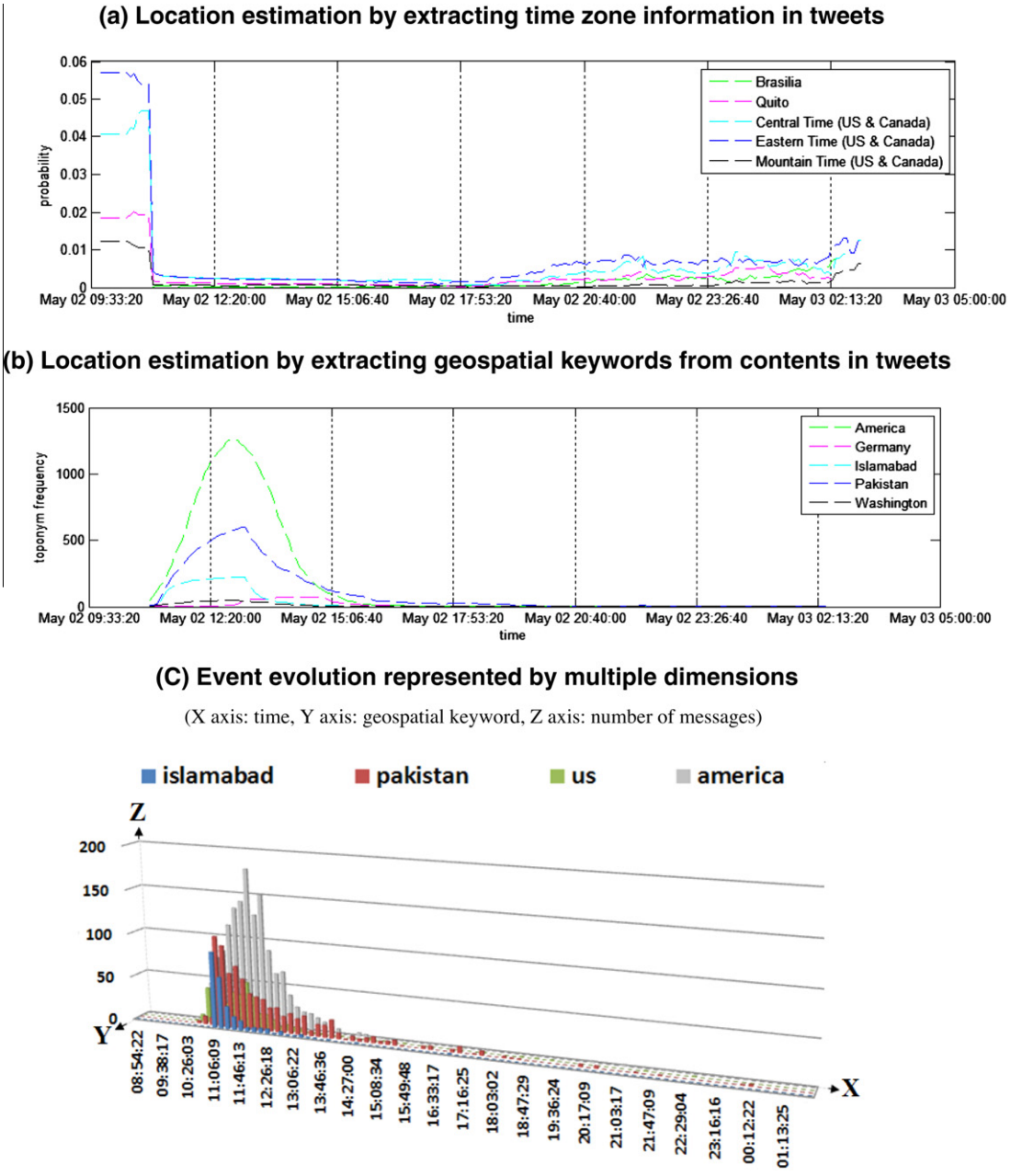


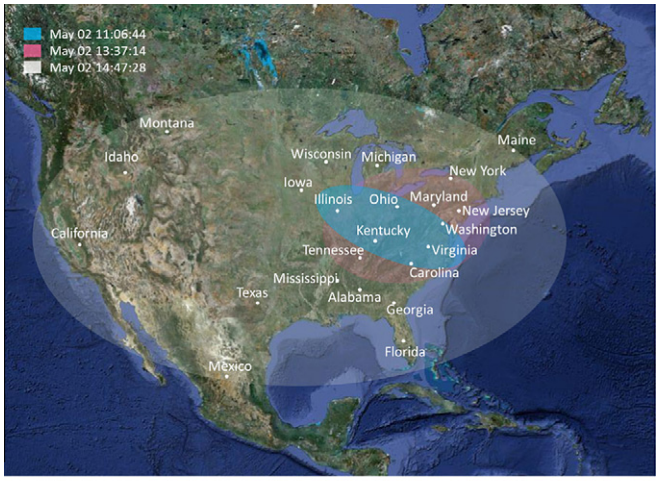
Fig. 18. (a) Event 3: Location estimation by extracting time zone information in tweets. (b) Event 3: Location estimation by extracting geospatial keywords from contents in tweets. (c) Event evolution representation based upon three dimensions.

discussion on death of Osama Bin Laden in Twitter have been suddenly raised to 100 time zones around the world, and the news therefore became a global event.

- Discussion (Experimental event III)
  (1) In this case, although the original event "death of Osama Bin Laden" was occurred in Islamabad, Pakistan, the factors that the place of Obama's presentation and this news were concerned by most American people make the related tweets were largely located in the geographical areas of the United States. However, the keywords "Pakistan" and "Islamabad" were frequently appeared in the headline titles of breaking news in some media (e.g. CNN breaking news) at the time point, so it allows for people only use the content-centric method, without the implication of time zone and geo-tagging information to find the original location of the event.
  (2) As shown in Fig. 17, the tweets related the event reached a peak point at 12:35, totally 28,659 tweets involved in the event discussion. The curve in the event timeline represents a typical power law distribution. Initially the number of related messages was quickly increased to reach a peak point, and subsequently the curve exponentially dropped from the top point, and then it gently went down to reach a low point. It is worth mentioning that our system is mainly used to detect real-time bursty events, and not concentrating on event tracking functions. In this case, the fact the topic cluster faded away shortly doesn't mean that no more tweets discussed the event topics from now on. Instead, it represents the event discussion became not as popular as before.
  (3) Differing from previous event cases, this event is not an unexpected natural disaster. The trajectory of event location in this case is an interesting issue. The event starts with the talk of President Obama. Hence it seems that the system cannot directly detect the actual location associated with the news story "death of Osama Bin Laden" at the first time point, although our system did obtain the location information of the story (i.e. Islamabad and Pakistan) eventually. As demonstrated in the experimental results, the topic drift detected from the Twitter messages is obvious. It seems to imply the news story behind the content of presentation of President Obama is more critical than the original event (i.e. Obama's talk). In addition, this case also suggests that the trajectory of event location in tweets is largely influenced by other news media (e.g. CNN breaking news). The messages related to the event have been mapped into the Google Map for visualization, as illustrated in Fig. 19.

6. Related work

Research on microblog has been popular in areas of text mining and information retrieval in recent years, the brand-new type of user-generated content challenges researchers to mine treasure from massive message streams. One of the key advantages of microblog is that it enables people to achieve a near real-time information awareness. Users in microblogs can be seen as social sensors (Rosi et al., 2011), the distributed sensors share their daily activities and information that may benefit us to understand "what's going on" in the world. For instance, (Sakaki, Okazaki, & Matsuo, 2010) utilized real-time characteristic of Twitter to monitor a set of specific keywords such as "earthquake", thus to detect earthquakes in Japan. They also considered each user as a sensor and applied Kalman filter and particle filter for location estimation. Utilize keywords to discovery topics is a common approach which is applied in many literatures and web services, such as Twitter-Monitor (Mathioudakis & Koudas, 2010)is a trend detection system which treats bursty keywords as entry points, bursty keywords



(a) Event visualization(event 3(i))

(b) Event visualization(event 3(ii))

Fig. 19. (a) Event visualization (event 3(i)), and (b) event visualization (event 3(ii)).

detection model and trend analysis model were designed to monitor events on Twitter. To gain better insight into trending topics, Cheong and Lee (2009) presented an excellent view of trending topics on microblogs. The topics on Twitter were categorized into short-term, medium-term and long-term topics. And another work also applied statistics on different aspects of trending topics. Naaman, Becker, and Gravano (2011) developed a location-centre method to categorize and characterize Twitter trends for some specific geographic area (e.g. New York City), and showed that not all trends are created equal. Their approach allowed for automatically distinguishing between different types of trends by analyzing Twitter dataset. Differing from our approach, the development of their method concentrated on handling Twitter messages from one geographic area, rather than utilizing the messages produced by the occurrence of some event in any place for analysis of Twitter trends.

On the other hand, detecting topics and analyzing trends on Twitter change the way people acquire information, but it still cannot fully satisfy user's information needs. People want to know where the event happened, and where these messages were sent. On August 2009, Twitter released the location service that enables mobile users publish their tweets with latitude and longitude[1]. Through this service, we can annotate real world map with messages. This process is so-called *geo-tagging*[2]. Thus, more and more

[1] "Location, Location, Location," http://blog.twitter.com/2009/08/location-location-location.html (August 20, 2009)

[2] "Twitter Tip: Geo-tagging. What is it, how to do it, and for God's sake, "Why?"," http://www.businessesgrow.com/2010/01/19/twitter-tip-geo-tagging-what-is-it-how-to-do-it-and-for-gods-sake-why/ (January 19, 2010).

scholars started to utilize location information on microblogging posts. One of related research is (Longueville, Smith, & Luraschi, 2009), this work analyzed the forest fire event at the area of southern France. With the analysis of temporal dynamics and spatial information, more and more relevant discussions were joining while the expansion of disaster area and time lapsing. One of the advantages of geo-tagging events is that it can help us realize the affected area.

In our survey, there were three main methods of acquiring spatial information to deal with spatio-temporal analysis from Twitter in the previous work. They are discussed as follows:

- *Latitude and longitude*: The most intuiting method is to annotate them with geographic coordinates. Lee and Sumiya (2010), Lee, Wakamiya, and Sumiya (2011) proposed a geo-social event detector which utilizes local crowd behaviors to detect unusual geo-social events. This method is based on a precise form of location (i.e. latitude and longitude). If there was a densely populated region, the processes of ROIs and geographical regularities will be used to inspect whether it is an unusual geo-social event.
- *Time zone*: Alternatively, the geographic information on tweets can also be obtained from user profiles. Some users may specify their location by typing the location field and selecting a time zone. Hecht, Hong, Suh, and Chi (2011) reveals that only 66% users on Twitter enter valid geographic information in the block of location field, because of they can write any types of string even a name of a pop star (someone located in "Justin Biebers heart"). This work also tried to apply machine learning to determine whether we can identify users' location from their posts. Some research teams utilized the time-zone based location information of the microbloggers who have posted micriblogs to analyze spatio-temporal information from tweet topics. For instance, Song, Li and Zheng (2010) presents a framework to mine the associations among topic trends in Twitter for related topic search by the extraction of topics' spatio-temporal information and the calculation of the similarity among topics. Cheong (2009) addresses statistics about aggregated messages with different cities; the result shows that a topic may be discussed by several regions.
- *Content*: Sankaranarayanan, Samet, Teitler, Lieberman and Sperling (2009) presented a system so-called TwitterStand which captures tweets that correspond to breaking news. In order to cooperating geographic information into events, the *part-of-speech* and *named-entity recognition* approach are used to extract geographic nouns from contextual posts. Cheng, Caverlee and Lee (2010) also proposed a method to automatically identify location keywords and further geo-locate users from message content. But location ambiguity is still a problem for extracting exact location information from content, which was also mentioned in Cheng's work.

To sum up, utilization of latitude and longitude data may be the most precise way for locating messages. However, according to our preliminary statistics on 270,852 sample tweets, we found that approximately 66,565(24%) tweets have no time-zone information from their user profile, and 268,831(99%) tweets have no latitude and longitude information through Twitter Stream API. As a result, the way of latitude and longitude information in tweets is not well suited for detecting popular geographical events in our work.

In contrast, extracting location keywords from text content seems to be the blurriest way. However, in some cases the way of extracting geospatial keywords from contents in tweets may be more reliable for location estimation. This is due to lots of users did not provide real location information in user profiles, which can mislead most geographic information systems. This situation has been identified in our experiments.

Another problem is that it is difficult to distinguish whether it is come from native speakers. Therefore, in this work we applied text mining to develop online topic detection system. Instead of using keywords to track events, we established a location analysis model to dynamically estimate where the event was happened.

Besides the few On the other hand, the clustering data stream (Khy, Ishikawa, & Kitagawa, 2008) has been an important issue in data mining community in recent years. STREAM (Guha, Meyerson, Mishra, Motwani, & O'Callaghan, 2003) is the firstly data stream clustering technique proposed by Guha et al. The *k*-median clustering algorithm was adopted with a simple algorithm based on divide-and-conquer to solve the space limitation problem. In addition, CluStream (Aggarwal, Han, Wang, & Yu, 2003) is also a stream clustering process that generates an online component which periodically stores detailed summary statistics and an offline component which uses only this summary statistics. Zhong (2005a) combined online spherical *k*-means (OSKM) algorithm with an existing scalable clustering strategy to achieve fast and adaptive clustering of text streams. In order to deal with online processing data with temporal information, forgetting (half-life[3]) mechanism has been utilized in lots of research work (Ishikawa, Chen, & Kitagawa, 2001; Uejima, Miura, & Shioya, 2004; Zhong, 2005a, 2005b) for decaying the cluster importance exponentially.

Experiments show that two online clustering algorithms OCTS (stands for Online Clustering of Text Streams) and OCTSM (stands for Online Clustering of Text Streams with Merge) have an almost satisfactory results in clustering quality, runtime and memory cost. Compared with their work, we use a similarity-based clustering approach instead of the model-based clustering method, so the half-life mechanism is not applicable in this case.

Generally speaking, in our survey most text stream clustering work use *k*-means techniques as their major text stream clustering algorithm. The main drawback of the *k*-means clustering method is that it should determine the fixed parameter of *k* (i.e. topic), and it is thus unsuitable for some real world applications in the problem domain, especially in dealing with the topic detection task with dynamic topics. Such issues were discussed in Zhong (2005a) and Roxy and Toshniwal (2009), and some solutions for avoiding empty cluster problems and choice k were addressed. Due to the problems of *k*-means clustering methods, we use online density-based approach for extracting topics from microblogging data collection.

7. Conclusion

Severe natural disasters such as earthquakes require new scientific methodologies for risk management and control. Understanding their possible impacts and striving towards their timely detection and prevention can help protect lives and properties. Through analyzing the temporal and spatial dynamics of Twitter activity, in this work we developed several algorithms for mining microblogging text streams to obtain real-time and geospatial event information. The goal of our approach is to effectively detecting and grouping emerging events by utilizing real-time messages and geolocation data provided by Twitter services. The preliminary results show that our platform model has the potential for event detection and awareness.

To the end, the conclusions of this work are listed as follows:

- In this work we have applied an online clustering approach for detecting emerging events and analyzing spatio-temporal information associated with the discovered events on Twitter

[3] Half-life is the period of time it takes for a substance undergoing decay to decrease by half.-from Wikipedia

messages, in order to enhance the understanding of the evolving events. This suggests that microblogs can be a deployable tool for situational awareness of unexpected sudden events.

- The aim of our approach is to offer a way to organize real-time event topics, allowing users to quickly figure out emerging events in the world. Compared with most commercial real-time search services, our method does not require any form of queries from users to fulfill information acquisition.
- Once the concepts behind the messages evolve with time, the underlying cluster structures may also significantly change with time. In the experiments, we have demonstrated that our system is able to adapt itself to support "concept drifting" in the microblogging text streams.
- The length of each message leads to a problem with the lack of semantic integrality in tweets. This makes it more difficult to design a reliable weighting and clustering algorithms. In this work, we overcome such challenges, by utilizing our developed dynamic weighting method. The preliminary results show that our algorithmic model has the potential for event detection and awareness.
- In addition to content-based techniques and our location approximation methods, part of the analysis of spatial areas in this work is based on time zone. In the experiments, the examples and the results we have demonstrated are based on large-scale events. Going further, we will improve our method to analyze events which are happened in small spatial areas in future work. In our experiments, we found that the time-zone-based approach appears to be insufficient to deal with such cases.

In our future work, we will mainly focus on three tasks on the topic. The first task is to conduct a detailed study on evaluating other candidate on-line clustering methods for fulfilling microblogging text mining, compared their performance with our developed density-based methods. On the other hand, some tweets do contain time zone and latitude/ longitude data, which could be used to build some ground truth data against which the experimental results from text mining could have been compared in order to evaluate the performance (mainly in terms of accuracy) of the proposed method. However, so far in our work no attempt to evaluate the method beyond face validity was done, even though it would have been possible. Thus, the second task is to study on better ways for mining event locations and developing evaluation methods for geospatial prediction of events. In the third task, the utilization of geospatial name-entity recognition (NER) techniques would be helpful for location estimation, and could be incorporated with our approach for identifying geo-location information in the microblogging text mining process.

References

Aggarwal, C.C., Han, J., Wang, J., & Yu, P.S. (2003). A framework for clustering evolving data streams. In *Proceedings of the 29th international conference on very large data bases,* Volume 29, Berlin, Germany.

Bifet, A. (2010). Adaptive Stream Mining: Pattern Learning and Mining from Evolving Data Streams. In *Proceedings of the 2010 conference on adaptive stream mining: Pattern learning and mining from evolving data streams*.

Brants, T., Chen, F., & Farahat, A. (2003). A System for New Event Detection. In *Proceedings of the 26th annual international acm sigir conference on research and development in information retrieval,* pp. 330–337, ACM, Toronto.

Bun, K. K., & Ishizuka, M. (2002). Topic Extraction from News Archive Using TFPDF Algorithm. In *Proceedings of the 3rd International Conference on Web Information Systems Engineering,* pp. 73–82, IEEE Computer Society, Los Alamitos.

Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management,* pp. 759–768, Toronto, ON, Canada.

Cheong, M. (2009). What are you Tweeting about?: A survey of trending topics within twitter. Clayton School of Information Technology (2009).

Cheong, M., & Lee, V. (2009). Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. In *Proceedings of the 2nd ACM workshop on Social web search and mining, Hong Kong, China*.

Ester, M., Kriegel, H. -P., Sander, J., Wimmer, M., & Xu, X. (1998). Incremental Clustering for Mining in a Data Warehousing Environment. In *Proceedings of the 24rd international conference on very large data, bases*.

Finch, T. (2009). Incremental calculation of weighted mean and variance. University of Cambridge.

Guha, S., Meyerson, A., Mishra, N., Motwani, R., & O'Callaghan, L. (2003). Clustering data streams: Theory and practice. *Proceedings of the IEEE Transactions on Knowledge and Data Engineering, 15*(3), 515–528.

Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the 2011 annual conference on Human factors in computing systems, Vancouver, BC, Canada*.

Ishikawa, Y., Chen, Y., & Kitagawa, H. (2001). An On-Line Document Clustering Method Based on Forgetting Factors. In *Proceedings of the fifth European conference on research and advanced technology for digital libraries*.

Khalilian, M., & Mustapha, N. (2010). Data Stream Clustering: Challenges and Issues. In *Proceedings of the international multiconference of engineers and computer scientists, Hong Kong*.

Khy, S., Ishikawa, Y., & Kitagawa, H. (2008). A novelty-based clustering method for on-line documents. *Proceedings of the World Wide Web, 11*(1), 1–37.

Lee, R., & Sumiya, K. (2010). Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks, San Jose, California*.

Lee, C.H., Chien, T.F., & Yang, H.C. (2010). DBHTE: A Novel Algorithm for Extracting Real-time Microblogging Topics. In *Proceedings of the 23rd international conference on computer applications in industry and engineering (CAINE 2010), Las Vegas, USA*.

Lee, C. H., Wu, C. H., & Chien, T. F. (2011). BursT: A Dynamic Term Weighting Scheme for Mining Microblogging Messages. In *Proceedings of the eighth international symposium in neural networks (ISNN 2011)*, Vol. 6677, Lecture Notes in Computer Science, pp. 548–557, Springer Berlin/Heidelberg.

Lee, R., Wakamiya, S., & Sumiya, K. (2011). Discovery of unusual regional social activities using geo-tagged microblogs. *Proceedings of the World Wide Web, 14*(4), 321–349.

Longueville, B. D., Smith, R. S., & Luraschi, G. (2009). OMG, from here, I can see the flames!: A use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 International Workshop on Location Based Social Networks, Seattle, Washington*.

Mathioudakis, M., & Koudas, N. (2010). TwitterMonitor: Trend Detection over the Twitter Stream. In *Proceedings of the 2010 ACM international conference on management of data,* pp. 1155–1158, Indianapolis, Indiana, USA.

Naaman, M., Becker, H., & Gravano, L. (2011). Hip and trendy: Characterizing emerging trends on Twitter. *Journal of the American Society for Information Science and Technology, 62*(5), 902–918.

Nguyen-Hoang, T. -A., Hoang, K., Bui-Thi, D., & Nguyen, A. -T. (2009). Incremental Document Clustering Based on Graph Model. In *Proceedings of the fifth ACM international conference on advanced data mining and applications, Beijing, China*.

Rosi, A., Dobson, S., Mamei, M., Stevenson, G., Ye, E., & Zambonelli, F. (2011). Social Sensors and Pervasive Services: Approaches and Perspectives. In *Proceedings of the 2011 IEEE International on Pervasive Computing and Communications Workshops (PERCOM Workshops)*.

Roxy, P., & Toshniwal, D. (2009). Clustering Unstructured Text Documents Using Fading Function. In *Proceedings of the World Academy of Science, Engineering and Technology,* pp. 149–156.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web, Raleigh, North Carolina, USA*.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., & Sperling, J. (2009). TwitterStand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems, Seattle, Washington*.

Song, S., Li, Q., & Zheng, N. (2010). A spatio-temporal framework for related topic search in micro-blogging. In *Proceedings of the 6th international conference on active media technology, Toronto, Canada*.

Sun, A., & Hu, M. (2011). Query-Guided Event Detection from News and Blog Streams. In *Proceedings of the, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, pp. 834–839.

Uejima, H., Miura, T., & Shioya, I. (2004). Giving Temporal Order to News Corpus. In *Proceedings of the 16th IEEE international conference on tools with, artificial intelligence*.

Wen, J.-R., Nie, J.-Y., & Zhang, H.-J. (2002). Query clustering using user logs. *Proceedings of the ACM Transactions Information System, 20*(1), 59–81.

Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning, 23*(1), 69–101.

Zhong, S. (2005a). Efficient streaming text clustering. *Neural Networks, 18*(5–6), 790–798.

Zhong, S. (2005). Efficient online spherical *k*-means clustering. In *Proceedings of the 2005 IEEE international joint conference on, neural networks*, pp. 3180–3185.