Expert Systems with Applications 39 (2012) 13338–13356

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Unsupervised and supervised learning to evaluate event relatedness based on content mining from social-media streams

Chung-Hong Lee *

Dept. of Electrical Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan

ARTICLE INFO

Keywords:
Stream mining
Data mining
Event evaluation
Social networks

ABSTRACT

Due to the explosive growth of social-media applications, enhancing event-awareness by social mining has become extremely important. The contents of microblogs preserve valuable information associated with past disastrous events and stories. To learn the experiences from past events for tackling emerging real-world events, in this work we utilize the social-media messages to characterize real-world events through mining their contents and extracting essential features for relatedness analysis. On one hand, we established an online clustering approach on Twitter microblogs for detecting emerging events, and meanwhile we performed event relatedness evaluation using an unsupervised clustering approach. On the other hand, we developed a supervised learning model to create extensible measure metrics for offline evaluation of event relatedness. By means of supervised learning, our developed measure metrics are able to compute relatedness of various historical events, allowing the event impacts on specified domains to be quantitatively measured for event comparison. By combining the strengths of both methods, the experimental results showed that the combined framework in our system is sensible for discovering more unknown knowledge about event impacts and enhancing event awareness.

1. Introduction

Event evaluation using social streams is a challenging area of research that attempts to evaluate evolving real-world events by utilizing continuously arriving message streams. The challenges normally come from the process of incremental clustering of unpredictable volume of incoming event elements in the dynamic environment. In most cases, the internal structures of news events in real world are quite complicated. How and why things in the events are regarded similar and related thus has deep-rooted consequences to how the model works. Given these conditions we still require effective methods by which to compare current and past events, and learn past experiences to cope with possible event evolution. Recently, due to the continuous growing presence of social-media applications, there has been a numerous research effort on developing solutions to employ social-media power for awareness of real-world events. Among these applications, one of the most significant phenomena is that people who are located close to the location of occurrence of some emerging event have a higher probability for reporting up-to-date situation about the most recent event development. Meanwhile, people lived in other countries concerning the same event may also contribute their insightful ideas regarding side-effects of event development through social networks. This provides useful knowledge for resolving problems once people suffer from similar events. While this pattern holds across a wide range of real-world cases and time periods, little attention has been paid to establish effective methods for evaluating event relatedness through the use of social media. In fact, the contents of microblogs preserve valuable information associated with past disastrous events and stories. To learn the experiences from past microblogging messages for coping with emerging real-world events, allowing make sensible decisions, the techniques for event evaluation are essentially required. Due to emerging real-world events continually evolve, it is hard to figure out the structure and dynamic development of emerging events, and directly utilize the data of the on-going event to compare with the ones of past events. Novel online event detection techniques, which corporate streaming models with online clustering algorithms, provide feasible solutions to deal with the text streams (e.g. Tweets) for event mining in real time. To estimate the impacts for event understanding, in this work we developed a framework of event evaluation system on Twitter dataset, and used the social-media messages to characterize the collected events for relatedness analysis.

This work is an attempt to describe the concept of event relatedness using social network datasets. We consider two aspects of relatedness computation we believe event relatedness model

* Tel.: +886 7 3814526.
E-mail address: leechung@mail.ee.kuas.edu.tw

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.05.068

should carry out. First, it should take the relatedness among the considered dimensions into account. Second, the relatedness measures should cover online and offline evaluation of detected events. By analyzing the contents of Twitter dataset, our work started with the formulation of event features. In our previous project, we have developed an online event detection system for mining Twitter streams using a density based clustering approach. In this work, we go on evaluating event relatedness using event clusters produced by the developed system platform. Some functions of the developed system framework have been reported in our previous work (Lee, 2012; Ester, Kriegel, Sander, Wimmer, & Xu, 1998; Lee, Wu, & Chien, 2011; Lee, Yang, Chien, & Wen, 2011).

The offline event-evaluation model emphasizes extensible capacity in the relatedness measure metrics. That is, the relatedness metrics should be able to be self-contained and also combine with the extensible amounts of specified event-topic elements for measures. These two models provide a unified framework for the relatedness analysis of current and past events, event association, and event evolution, etc.

We propose here a combined framework to establishing relatedness evaluation techniques that incorporates a model for online evaluation of emerging event and measure metrics for offline event evaluation. In this work, the results of relatedness measures were mainly based on a quantitative assessment of relatedness among events, which can be used to support analyzing the implicit relationships among events, providing insightful viewpoints for event awareness. This is a novel approach in this field by validating considered impact factors involved in the event development, for contributing to relatedness evaluation and analysis of real-world events.

1.1. Problem statements

What makes a past event-story related to the current event? Presumably, two event-stories should be contextually or conceptually related to each other. In this perspective, *similarity* would be an important representation of relatedness. However, similarity is not a sufficient attribute of the problem at hand. In previous work relatedness between two events is often represented by similarity between these events. In this problem domain, '*relatedness*', however, is a more general concept than '*similarity*'. Similar events are obviously related by virtue of their similarity, but dissimilar events may also be implicitly related by some other hidden relationships, although these two terms are used sometimes interchangeably. For the applications of event analysis, evaluation of relatedness is more helpful than similarity, since there are quit a lot of implicit and useful clues with dissimilar features among various events. Thus, in this work we established a novel combination of several techniques for evaluating events' relatedness, rather than only work on computing their similarity.

Given the fact that some past events were still better understood by past news documents, our goal of this work is, roughly speaking, to extract features from a tweet corpus for locating a list of related events that the user would like to study afterwards. Hence, through the developed system, the users will be able to estimate the likelihood that the equivalence relation holds for a given collection of event datasets based on their selected features.

2. System framework

2.1. Problem characteristics

In this section, the system framework and architecture for evaluating relatedness of detected events based upon Twitter data is described. First of all, we present some problem characteristics and difficulties in system development for exploring microblogging contents associated with the event relatedness as below.

- Alone with the online event detection functional module, the system framework also combines online- and offline-event evaluation subsystems. While these systems were able to be integrated together, the functions of online and offline evaluation subsystems are allowed to be stand-alone. That is because that we expect to create an opportunity for people to make judgments regarding the analysis of event impacts by different implementation methods and empirical results through unsupervised and supervised models for the application domain.
- For empirical purposes, an event representation should at least share some similar content with the ones of other related events. The notion of relevance in information retrieval, which measures to what extent the topic of a candidate document matches the topic of the query, can be regarded as a natural form of relatedness. A variety of retrieval models have been well studied in the field of information retrieval to model relevance, such as vector space model. Motivated by this, in this work it is expected to build effective models to measure and analyze intrinsic relatedness among events for meeting users' information needs.
- Exchanging microblogging messages depends on the users to communicate, which almost always needs that they are competent in the same language or rely on a bilingual mediator. It is clear that language stands in a more complicated relationship to the formulation of discussed events than national boundaries. In reality, many people currently around the world use English in addition to their local and national languages. For instance, most of the Twitter users located in the English-speaking countries follow users who are located in English-speaking countries. However, even for local and overseas communication of Twitter users located in non-English speaking countries, the use of the same dominant language (i.e. English) in discussions regarding specific event is still significant. This suggests that the effect of country-specific-language might be weakened for event analysis by the wide use of English as a *lingua franca*. The view is taken, therefore, in this work we focus only on the use of English as the major language for study of microblog based event evaluation.
- Once dealing with Twitter streams, one important factor is the presence of message locations. In particular, when analyzing the distribution of Twitter messages for event awareness, it is important to consider the uneven distribution of the users' locations around the world. Of course, Twitter users are certainly not distributed evenly around the world. However, the event evaluation method applied in our work is mainly concerned with automatic identification of *bursts* from Twitter posted messages, providing useful insights into the local events and in turn facilitating timely event monitoring. This may benefit the study of the first hand information from the original locations of the event occurrence. However, for long-term analysis of event development, we still need to take such an issue into account.
- In reality, the concept transition related to evolving events is generally hard to be estimated. It would be quite difficult to analyze transitions of unknown events. In particular, extraction of sensible information from a clustering result is not an easy task. Under such a circumstance, background knowledge regarding impacts gained from past events may be helpful. Thus, in this work we also look at several other dimensions that can either strengthen or impede the extent of relatedness between events. In addition to spatial and temporal analysis, we established an extensible measure metrics covering several factors such as *business*, *politics*, *sport*, *climate*, *commodity*,

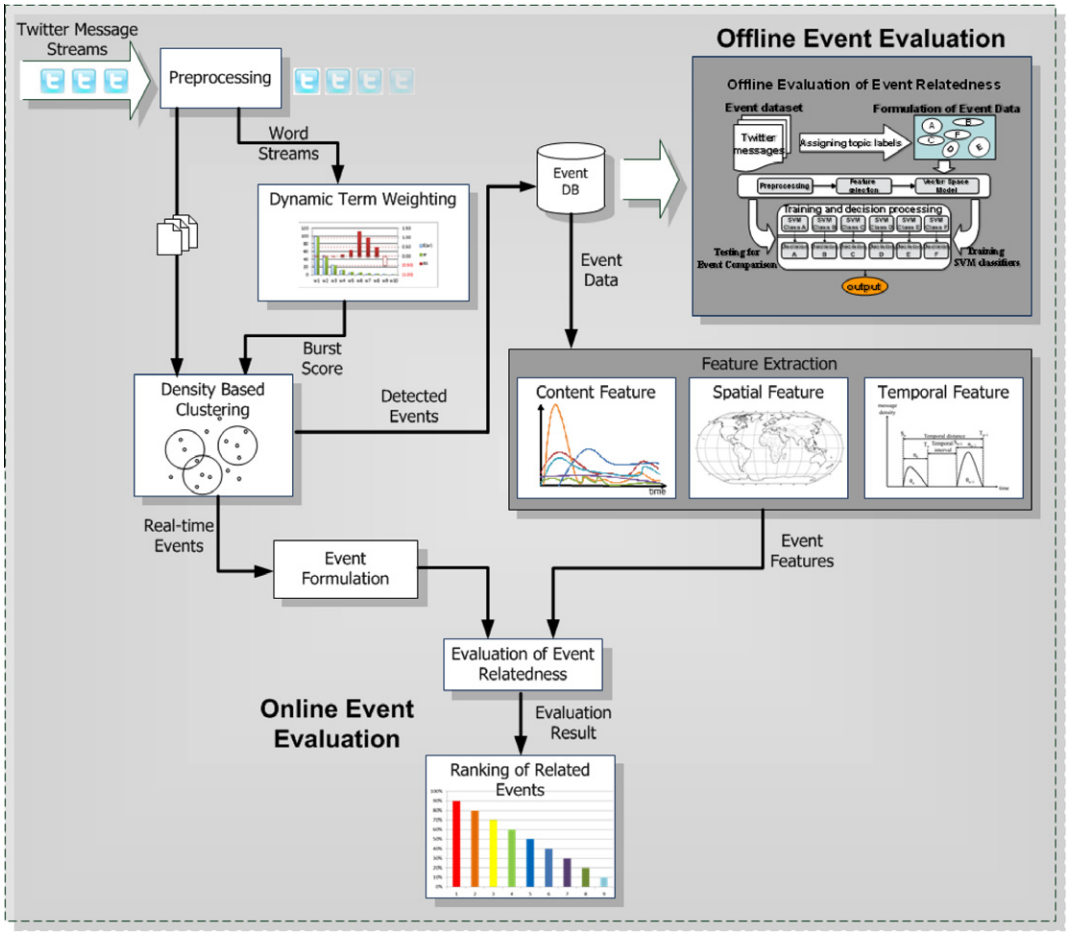


Fig. 1. The system framework.

finance, and *entertainment* to estimate the impacts of studied events. Part of the functions in the developed system provides a quantitative investigation of the effect of important factors for historical events to pursue a deep understanding of known events and their possible inter-relationship.

To solve the aforementioned issues, we have proposed a framework for event evaluation by mining contents from Twitter messages. We established an online evaluation approach on Twitter microblogs for detecting large-scale events and performing relatedness evaluation based upon an unsupervised technique. Furthermore, we have studied the development of measure metrics for offline evaluation of event relatedness. By supervised learning, our developed measure metrics are able to compute relatedness of historical events, allowing the event impacts on specific domains to be quantitatively analyzed, and evaluated to perform event comparison.

2.2. System architecture

In this section, the system framework and algorithms for mining events and evaluating relatedness based upon Twitter data (i.e. Tweets) is described. First, non-ASCII messages will be bypass for collecting ASCII coding content of messages. Subsequently, in order to perform unsupervised event clustering our system started with construction of a dynamic feature space which maintains messages with a sliding window model to deal with the message streams. New incoming messages will be reserved in memory till they are out of the window. Then we utilized a dynamic term weighting scheme (Lee, Wu, et al., 2011) to assign dynamic weights to each word. 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 density based clustering approach 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 by analyzing the collected cluster records. In order to measure the relatedness among events, we extract feature patterns of each event by performing content mining for content analysis, spatial analysis, and temporal analysis, as shown in Fig. 1. The datasets of detected events were stored in the event repository (i.e. Event DB). This allows our online approach to compare the new event vector with other event vectors for dynamic evaluation of event relatedness. On the other hand, the datasets of detected events stored in the repository are being used to perform offline relatedness measures and analysis by the developed supervised event evaluation method. More detailed description of our proposed approaches will be addressed later.

3. Characterization of detected real-world events

To further describing the event formulation, an example of detected event (i.e., "Japan earthquake on March 11, 2011") using Google Maps for illustration of geographical locations of the event is shown in Fig. 2. Fig. 3 illustrates the event evolution representation based upon different factors, including *time*, *geospatial*

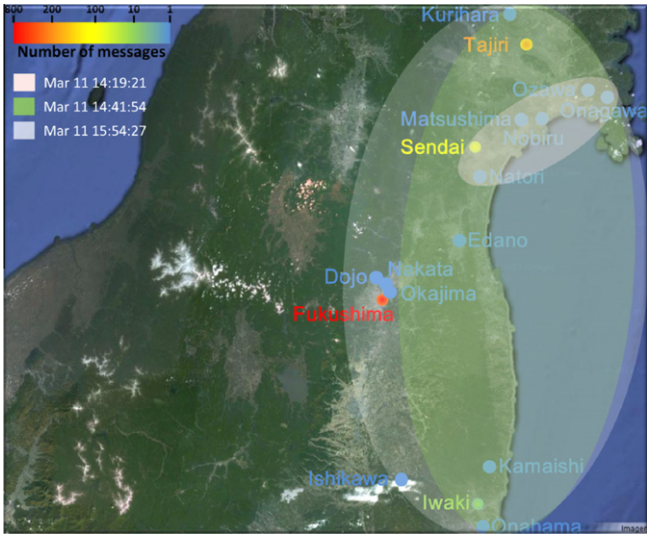


Fig. 2. The illustration of geographical locations for Japan earthquake event (March 11, 2011).

keyword, and the logarithm of the *number of messages*. The event timeline is often utilized to report the tweet activity by volume. Fig. 4 illustrates the sample Twitter-messages for Japan earthquake (March 11, 2011).

3.1. Content mining for extracting event's content feature

There are millions of short messages containing several keywords in Twitter service every day. The importance of these keywords will change over time. 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 (Becker, Naaman, & Gravano, 2010). Thus, in this work for mining Twitter message streams, we utilized a dynamic term weighting scheme called *BursT* (Lee, Wu, et al., 2011) to timely update the weighting of keywords in each messages. Subsequently, each message will be clustered by IncrementalDBSCAN clustering algorithm (Ester et al., 1998), and then our system will record the maximum burst weighting of each keywords of each cluster.

3.2. Content mining for extracting event' temporal feature

In our event evaluation system, we assumed that each event 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 keywords tracking methods is due to that such techniques can group relevant posts based on similarity of messages, avoiding missing valuable 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. 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. To process incoming texts with a chronological order, a fundamental issue we concerned is how to find the significant features in text streams. Besides, 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.

3.3. Content mining for extracting event's spatial feature

While an event occurs in real world, the Twitter users post messages which may contain spatial information regarding the event situation. Thus, by extracting the geographical terms from the content of these messages, the spatial information about where the event originally occurred and diffused can be obtained. We utilized GeoName which is a geographical dictionary to extract geographical terms form each clusters, and utilized *term frequency* weighting factor to weight each geographical terms for representing each

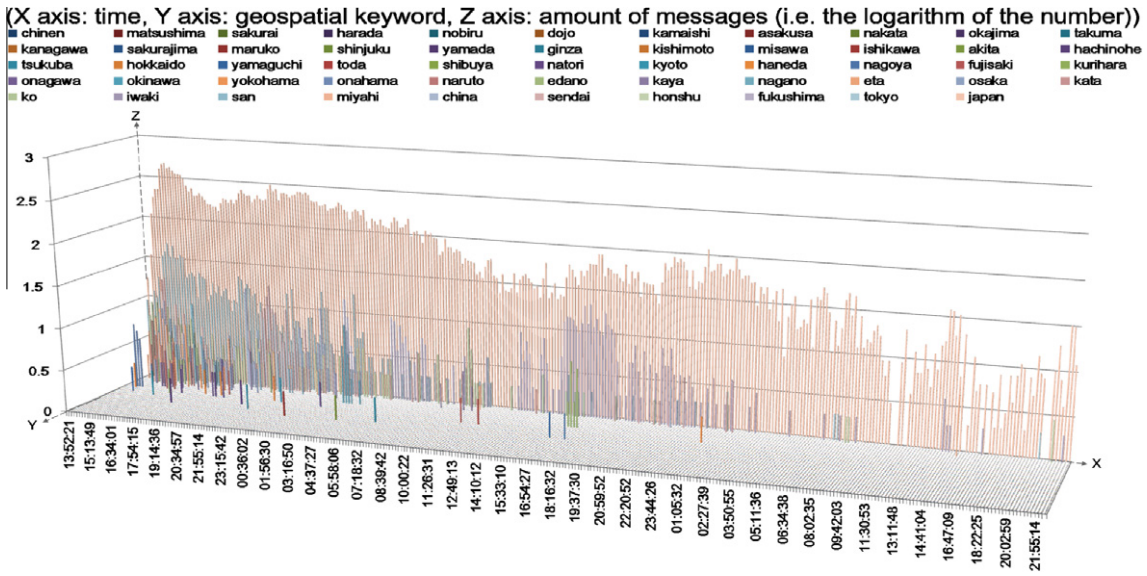
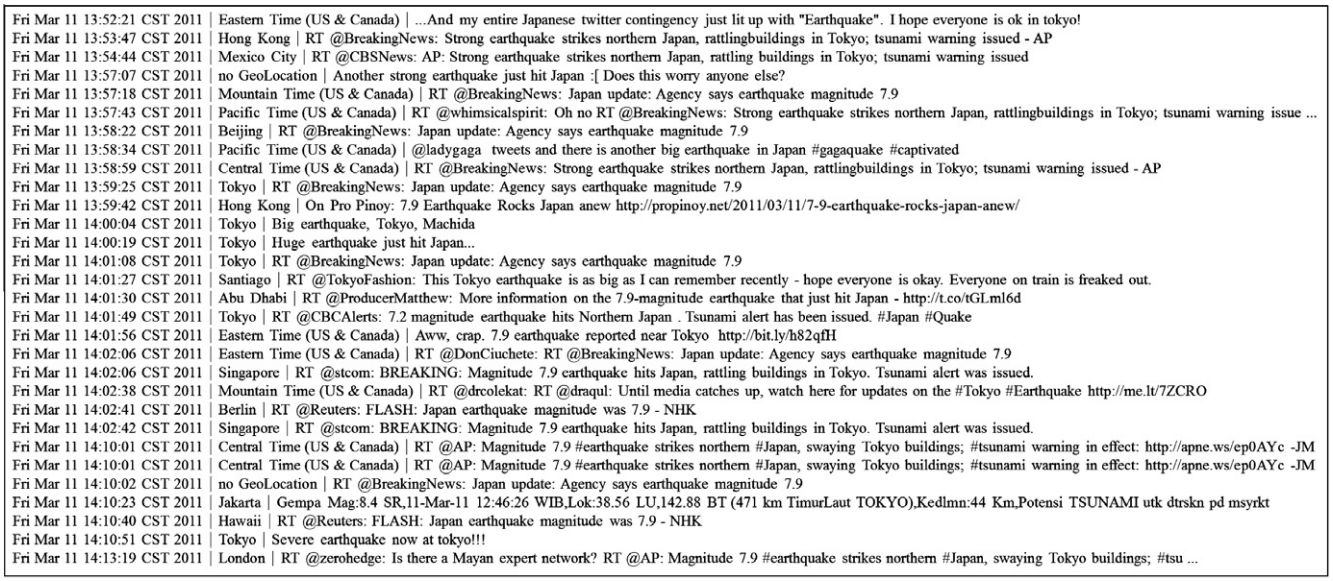


Fig. 3. The timeline of sample Twitter-messages for Japan earthquake (March 11, 2011).



Fri Mar 11 13:52:21 CST 2011 | Eastern Time (US & Canada) | ...And my entire Japanese twitter contingency just lit up with "Earthquake". I hope everyone is ok in tokyo!
Fri Mar 11 13:53:47 CST 2011 | Hong Kong | RT @BreakingNews: Strong earthquake strikes northern Japan, rattlingbuildings in Tokyo; tsunami warning issued - AP
Fri Mar 11 13:54:44 CST 2011 | Mexico City | RT @CBSNews: AP: Strong earthquake strikes northern Japan, rattling buildings in Tokyo; tsunami warning issued
Fri Mar 11 13:57:07 CST 2011 | no GeoLocation | Another strong earthquake just hit Japan :[Does this worry anyone else?
Fri Mar 11 13:57:18 CST 2011 | Mountain Time (US & Canada) | RT @BreakingNews: Japan update: Agency says earthquake magnitude 7.9
Fri Mar 11 13:57:43 CST 2011 | Pacific Time (US & Canada) | RT @whimsicalspirit: Oh no RT @BreakingNews: Strong earthquake strikes northern Japan, rattlingbuildings in Tokyo; tsunami warning issue ...
Fri Mar 11 13:58:22 CST 2011 | Beijing | RT @BreakingNews: Japan update: Agency says earthquake magnitude 7.9
Fri Mar 11 13:58:34 CST 2011 | Pacific Time (US & Canada) | @ladygaga tweets and there is another big earthquake in Japan #gagaquake #captivated
Fri Mar 11 13:58:59 CST 2011 | Central Time (US & Canada) | RT @BreakingNews: Strong earthquake strikes northern Japan, rattlingbuildings in Tokyo; tsunami warning issued - AP
Fri Mar 11 13:59:25 CST 2011 | Tokyo | RT @BreakingNews: Japan update: Agency says earthquake magnitude 7.9
Fri Mar 11 13:59:42 CST 2011 | Hong Kong | On Pro Pinoy: 7.9 Earthquake Rocks Japan anew http://propinoy.net/2011/03/11/7-9-earthquake-rocks-japan-anew/
Fri Mar 11 14:00:04 CST 2011 | Tokyo | Big earthquake, Tokyo, Machida
Fri Mar 11 14:00:19 CST 2011 | Tokyo | Huge earthquake just hit Japan...
Fri Mar 11 14:01:08 CST 2011 | Tokyo | RT @BreakingNews: Japan update: Agency says earthquake magnitude 7.9
Fri Mar 11 14:01:27 CST 2011 | Santiago | RT @TokyoFashion: This Tokyo earthquake is as big as I can remember recently - hope everyone is okay. Everyone on train is freaked out.
Fri Mar 11 14:01:30 CST 2011 | Abu Dhabi | RT @ProducerMatthew: More information on the 7.9-magnitude earthquake that just hit Japan - http://t.co/tGLml6d
Fri Mar 11 14:01:49 CST 2011 | Tokyo | RT @CBCAlerts: 7.2 magnitude earthquake hits Northern Japan . Tsunami alert has been issued. #Japan #Quake
Fri Mar 11 14:01:56 CST 2011 | Eastern Time (US & Canada) | Aww, crap. 7.9 earthquake reported near Tokyo http://bit.ly/h82qfH
Fri Mar 11 14:02:06 CST 2011 | Eastern Time (US & Canada) | RT @DonCiuchete: RT @BreakingNews: Japan update: Agency says earthquake magnitude 7.9
Fri Mar 11 14:02:06 CST 2011 | Singapore | RT @stcom: BREAKING: Magnitude 7.9 earthquake hits Japan, rattling buildings in Tokyo. Tsunami alert was issued.
Fri Mar 11 14:02:38 CST 2011 | Mountain Time (US & Canada) | RT @drcolekat: RT @draqul: Until media catches up, watch here for updates on the #Tokyo #Earthquake http://me.lt/7ZCRO
Fri Mar 11 14:02:41 CST 2011 | Berlin | RT @Reuters: FLASH: Japan earthquake magnitude was 7.9 - NHK
Fri Mar 11 14:02:42 CST 2011 | Singapore | RT @stcom: BREAKING: Magnitude 7.9 earthquake hits Japan, rattling buildings in Tokyo. Tsunami alert was issued.
Fri Mar 11 14:10:01 CST 2011 | Central Time (US & Canada) | RT @AP: Magnitude 7.9 #earthquake strikes northern #Japan, swaying Tokyo buildings; #tsunami warning in effect: http://apne.ws/ep0AYc -JM
Fri Mar 11 14:10:01 CST 2011 | Central Time (US & Canada) | RT @AP: Magnitude 7.9 #earthquake strikes northern #Japan, swaying Tokyo buildings; #tsunami warning in effect: http://apne.ws/ep0AYc -JM
Fri Mar 11 14:10:02 CST 2011 | no GeoLocation | RT @BreakingNews: Japan update: Agency says earthquake magnitude 7.9
Fri Mar 11 14:10:23 CST 2011 | Jakarta | Gempa Mag:8.4 SR,11-Mar-11 12:46:26 WIB,Lok:38.56 LU,142.88 BT (471 km TimurLaut TOKYO),Kedlmn:44 Km,Potensi TSUNAMI utk dtrskn pd msyrkt
Fri Mar 11 14:10:40 CST 2011 | Hawaii | RT @Reuters: FLASH: Japan earthquake magnitude was 7.9 - NHK
Fri Mar 11 14:10:51 CST 2011 | Tokyo | Severe earthquake now at tokyo!!!
Fri Mar 11 14:13:19 CST 2011 | London | RT @zerohedge: Is there a Mayan expert network? RT @AP: Magnitude 7.9 #earthquake strikes northern #Japan, swaying Tokyo buildings; #tsu ...

Fig. 4. Sample Twitter-messages for Japan earthquake (March 11, 2011).

clusters. Also, the spatial features of the event for location estimation can be obtained by extracting time zone data or based on a precise form of geographic coordinates of location (i.e. latitude and longitude) in tweets. However, according to our previous work a 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 geographical events in our work.

4. The unsupervised method for online text stream clustering and event evaluation

In this chapter, the developed online text-stream clustering approach for event evaluation by mining microblogging message streams is described. The developed online clustering method which is based on a real-time event-cluster generation model, including three parts: a dynamic term weighting scheme, a sliding window model, and an online density-based clustering approach.

Our work starts with developing the system for detecting topics and tracking events about hot news topics, and preferences of people from text information sources of Twitter 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. More details regarding system implementation can be found in our previous publication (Lee, Wu, et al., 2011).

4.1. Online text-stream clustering by a density-based 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.

4.1.1. Reasons for adopting a density-based approach on online event clustering

Adopting density-based clustering methods in this work is based upon the following reasons:sons:

- Density-based clustering techniques are capable of detecting arbitrary-shaped clusters.
- In microblogging messages, the contents normally include lots of noises. Once mining these messages, the clustering algorithm should be able 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.
- There is no assumption about the number of clusters with fixed topics, 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 et al., 1998) was used for our system development. IncrementalDBSCAN is an efficient algorithm which is based on DBSCAN for mining data with density-based connectively (Lee, 2012; Lee, Wu, et al., 2011; Lee, Yang, et al., 2011; (Ester et al., 1998). According to the theory of IncrementalDBSCAN clustering method, the shape of clusters will change over time when a

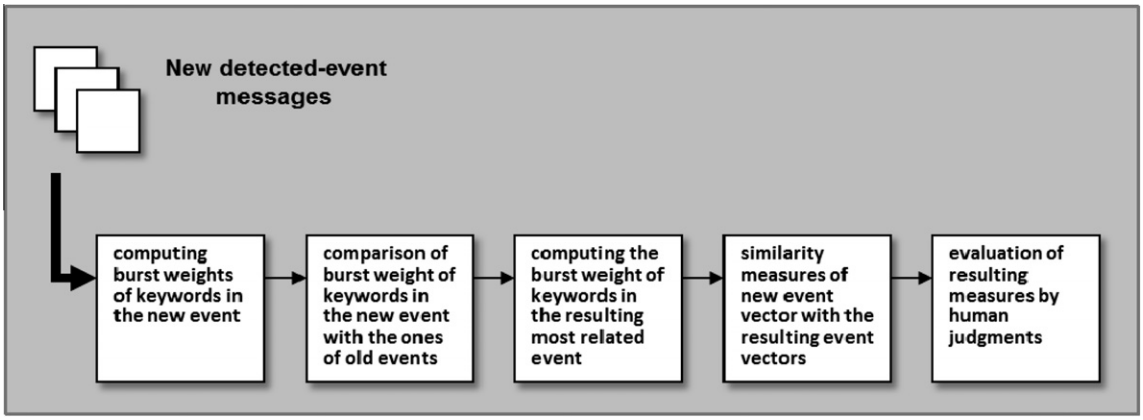


Fig. 5. The process of online event evaluation.

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.

4.2. A dynamic term weighting approach

It is a critical issue to find the significant features in text streams with a chronological order. In general information systems, weighting schemes use information that is based upon processing keyword distributions across the entire corpus. However, in this problem domain the text-stream corpora tend to be dynamic, with new messages always being added and weighting calculation values being updated. Thus, the significance of keywords in the text streams is always not stable but change with time. That is, the weighting values for microblogging message should be constantly changed. In particular, a special consideration is that almost all terms occur in each message only once due to the length limitation of microblogging messages. As a result, the computation overhead of *term frequency* (*tf*) would be strongly affected by the limited length. Besides, the *document frequency* (*df*) conflicts the operation of event-topic mining since a higher df value of the words implies the terms occur in many documents, which might lead to the problem of missing topic words in messages to some extent. As a result, we developed a special term weighting scheme called *BursT* which was proposed in our previous work (Lee, Wu, et al., 2011). The experimental results show that a better performance by utilizing our approach in weighting words of microblogging messages than many weighting methods (Lee, Wu, et al., 2011).

In this work, we used our developed weighting method (Lee, Wu, et al., 2011) which considers the characteristics of microblogs and incorporates burst detection for adapting dynamic environment. The solution of BursT weighting method is that a heavier weight is determined by a higher burstiness, in which some word occurs frequently in the window. Thus, the formula of BursT weighting scheme is shown in Eq. (1):

$$BursT_{w,t} = BS_{w,t} * TOP_{w,t} \tag{1}$$

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 $\langle w, \mathrm{atw}, t-1, \mathrm{nw}, t, \mathrm{E}(\mathrm{arw}, t)\rangle$, atw, $t-1$ represents the last time word w arrived, nw, t counts the total number of word w appeared in our system, and $\mathrm{E}(\mathrm{arw}, t)$ is a long time cumulative expectation of arrival rate to the word w.

The second factor 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 valid topic. Thus, the term occurrence probability corresponding to the word w at tth arrival is formulated as below:

$$TOP_{w,t} = P(w_t|c_t) = \frac{\{|m : w_t \in c_t\}|}{|c_t|} \tag{2}$$

where TOP represents the probability of the word occurrence in the sliding window, and ct denotes the message collection in the corpus collected from the time $t - \mathrm{tw}$ to current time. This factor would enable the weight of the word to grow with its occurrence frequency in messages, for identification of event topics (Lee, 2012).

4.3. Online generation of event clusters and dynamic relatedness evaluation

In our work, each extracted keyword in the tweet was assigned with burst weighting value for real-time event detection. Since the burst weighting value of each keyword for representing on-going event is dynamically changed over time, the system will keep the maximum burst weighting value of each keyword of the event for establishing an event-vector representation.

Once some emerging events were detected by our system, the event clusters and event vectors can be generated by formulating clustered messages. Also, a relatedness measure metrics developed for computing event relatedness is activated for event evaluation. Several essential features of each detected event dataset have been extracted for event formulation by performing content mining operations. This allows our approach compare the new event vector with other event vectors for evaluation of event relatedness.

Subsequently, we start to perform online relatedness measures among the on-going event and historical event vectors. For dynamic relatedness evaluation, we composed a new event vector by assigning updated burst weighting value, and then employed cosine similarity measure to calculate the vector relatedness among on-going event and historical events per ten minutes. The process of online event evaluation is illustrated in Fig. 5.

5. The supervised method for offline evaluation of event relatedness

Relatedness represents how well a candidate event is related to other events. In order to model relatedness, we propose several algorithmic evaluation methods that characterize relatedness from multiple aspects. As mentioned previously, we have established an

unsupervised online clustering approach to detect burstiness on Twitter microblogs for detecting realtime large-scale events, and performed online dynamic evaluation of event relatedness. Going further, our work moved to develop a measuring method for offline evaluation of event relatedness. Through our developed measure metrics for computing relatedness of historical events, the essential aspects of impacts of related events can be quantitatively evaluated and analyzed, allowing for working as a stand-alone system for event evaluation, or cooperating with the developed online event evaluation system for understanding possible event development and evolution.

5.1. Techniques for offline measures of event relatedness

Our base model structure is developed for comparing relatedness among event datasets from various perspectives. For impact analysis of events, the learned model developed in this work attempts to measure the extent of relatedness among several event by comparing their datasets in terms of several essential dimensions using a supervised classifier based metrics. As mentioned previously, through formulating the collection of related social messages an event-story can be modeled by a number of hidden topics and selected features, with each topic gathering a series of observed messages according to topic-specific terms and sentences used in the content. A representation of the event comparison can be done by the quantitative values represented in each defined topic-dimension in a classifier (topic)-based space in a supervised manner. This measures the model's descriptive power, while requiring no submitted on-going event data for comparison.

Thus, for offline event evaluation in this work we implemented a measure metrics for acquisition of event relatedness from Twitter messages by means of construction of a classifier-based system (i.e. *Support Vector Machines* classifiers). The *Support Vector Machines* (*SVM*) (Vapnik, 1999) is one of the major statistical learning models. It basically provides a way for data categorization by producing a decision surface to separate the training data samples into two classes. As such, the resulting classifiers are capable of discriminating similar/dissimilar (or related/unrelated) event data, and further computing the degree of relatedness among the event datasets by means of our developed algorithm. In this work, we utilized the *LIBSVM* (Chang & Lin, 2001) and a *RBF* kernel to evaluate our offline event relatedness.

5.2. Vectorizing event-data in a SVM-classifier based vector space

A decision combination function must make use of useful representations of classifier decisions. In Fig. 6, a SVM-based metrics system for measuring event relatedness is illustrated. For measuring event relatedness, in this work a new vector space was formulated by SVM-based classifiers as a measure metrics for measuring event represented by collected Twitter messages, in which the resulting decision values of the input event would be examined by each trained SVM classifier in the metrics. That is, event vectors in such a combination of classifiers were regarded as mappings of topic categories in points on a multiple dimensional grid to form a category vector space. This also reflects a real world situation that a single event may involve one or several categories of topics. The vector approach allows for a mathematic and a physical representation of events for measures of their relatedness in aspects of selected topics. Additional classifiers can be added to the model by adding another dimension to the geometric representation. The pattern of adding dimensions to represent additional classifiers can be continued as many times as needed. If we needed to model n distinct classifiers, then we would use such an extensible metrics (i.e., weight of classifier 1, weight of classifier 2, weight of classifier 3, ..., weight of classifier n) for vector evaluation. The classifier

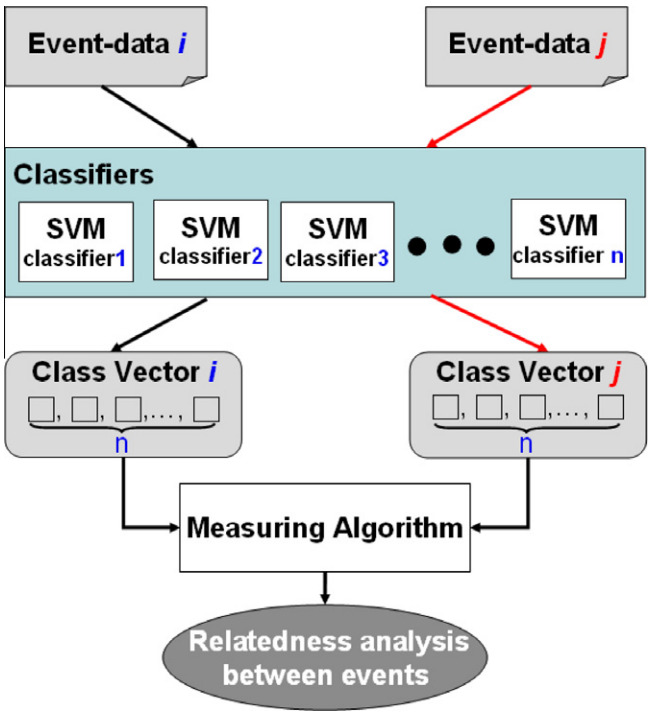


Fig. 6. A SVM-based metrics system for measuring event relatedness.

vector space uses the notion of a **space of topic category**, where each event is represented as a vector in a high-dimensional space. Because the position of an event in the vector space is determined by the degree of relevance based on the judgments of topic-classifiers, events with many topics in common end up close together, while events with few shared topics end up far apart. As a result, the relatedness between event vectors can be computed by means of several developed algorithms.

5.3. Metrics for measuring event relatedness using a multiple SVM-based classifier system

In this work we developed an approach applying a classifier-based technique with Support Vector Machines (SVM) method to support measuring of event relatedness. In the first stage, we employed topic specific messages to train Support Vector Machines (SVM) classifiers for constructing a measure metrics for identifying the topic categories of the events. Subsequently, we combined the trained classifiers to form a metrics, and input some unknown event data into the model to evaluate the decision values by each classifier in the metrics. Finally, new vectors that are formulated by resulting decision values from several different SVM classifiers (see Fig. 6) can be generated, and these vectors allowed us compute the extent of relatedness among events in a quantitative manner based on several measuring algorithm.

6. Experimental results and discussion

6.1. Experimenting with online event evaluation

In this work we experimented with a vast amount of Twitter data to identify the validity of the framework, through demonstrating selected cases by taking the events detected by the developed platform.

6.1.1. Case study (I): baseline event: "Virginia earthquake on August 24, 2011"

In the experiment, a total number of 192,541,656 Twitter posts were collected, dating from: January 1, 2011 to September 30, 2011. The test samples were collected through Twitter Stream API. After filtering out non-ASCII tweets, 102,709,809 tweets had been utilized as our data source. We utilized the dataset collected from January 1, 2011 to May 31, 2011 corpus as our dataset for training, and used the corpus dating from June 1, 2011 to September 30, 2011 as our test data. Subsequently, we partitioned messages into unigrams and all capital letters in each tweet were converted into lowercase for our experiments.

- **The datasets** (Case I: baseline event "Virginia earthquake on August 24, 2011").

In this experiment, we utilized the event "Virginia earthquake" as a baseline for identifying our framework. The event happened at 01:51, and the first post appeared at 01:52:04. The event was detected by our system at 01:52:17. The result of relatedness ranking

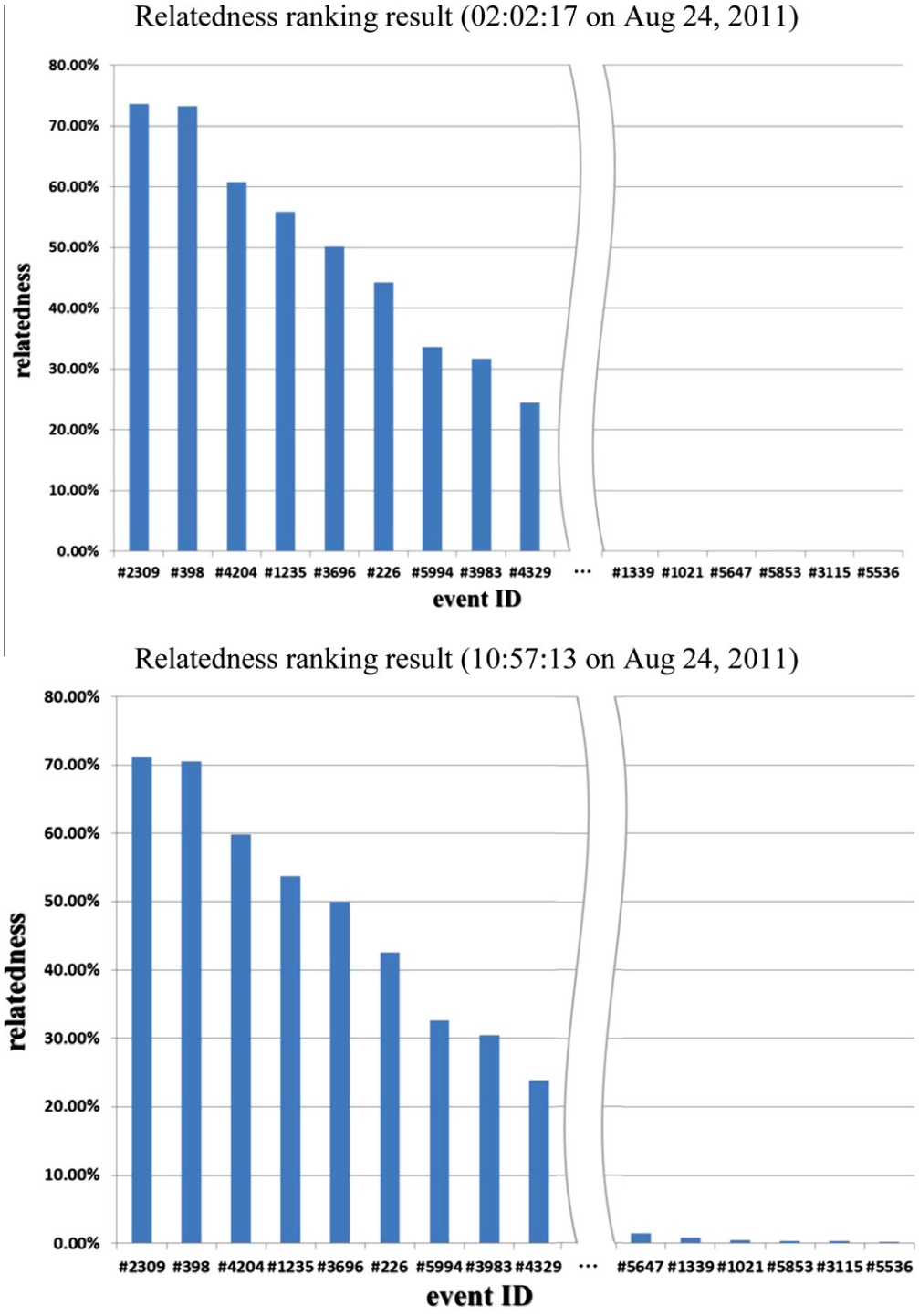


Fig. 7. Ranking of event relatedness upon a comparison with baseline "Virginia earthquake" event at various time points (Event ID: #1173, August 24, 2011).

Table 1
Top 10 keywords and the burst weighting in the detected events.

Event ID	Keywords and burst weighting
#1173	Earthquake:28.00598560037599, felt:4.2383560493654455, virginia:4.128908156333142, coast:3.369257348081017, shaking:2.976224432226175, dc:2.5043717231018507, east:1.7969757120599217, hurricane:1.5293515057891527, washington:1.081279754253795, experienced:0.915279825799065, nyc:0.8007819568139094, earth:0.6799292320901319, strikes:0.5157501676457266, damage:0.50535997014575, shake:0.48413468642430485, cnn:0.4388330326820492, building:0.4163997476679587, york:0.3994667787155791, quake:0.37886722363232483, 2012:0.3516273282981725
#2309	Earthquake:3.9553197961748126, zealand:0.963100917277774, prayers:0.909794403440777, 0800:0.884466396150691, affected:0.8383470931121912, nz:0.8314472900536037, cricket:0.7336453038423575, 65:0.6659161651840394, 22:0.6243486226423245, coverage:0.6216531727745904, coverage:0.6216531727745904, multiple:0.6140122444698878, england:0.5849805986667435, needing:0.5378653074347548, amb:0.46970540671599365, twibbon:0.4643744749876894, finder:0.4637452005497233, damage:0.4481274341139184, cellphone:0.4474914838300027, batteries:0.44564738622940314, urged:0.44522581056118227
#398	Earthquake:1.04908551454045, pakistan:0.8980141224744113, hits:0.16401090117859768, 18:0.15131980804529255, sw:0.12150740246713974, western:0.10616833459240625, major:0.09292551543064907, south:0.09152100337498355, means:0.08779825561221398, hit:0.08275405837041092, damage:0.06498796460480219, islamabad:0.06238490347410458, time:0.04667894881786543, today:0.04412245369384307, omg:0.036096448529406494, wow:0.03530149197516841, 40:0.035266668147545234, depth:0.03384105654844033, shallow:0.03294436042401517, news:0.03143891614025084
#4204	Philippines:0.49761490889948634, earthquake:0.47233581624246196, azkals:0.24853693404998034, safe:0.10928621389049406, manila:0.09546108516851091, birthday:0.08997353333089236, felt:0.08747401235101598, wnw:0.08061196314147687, 21:0.07613447410948651, magnitude:0.07176278026603079, stay:0.05649396739927458, 49:0.050996746536870474, tweet:0.05088502992889284, 57km:0.043150260103657585, hope:0.03410998649202747, shes:0.020935916385136415, huh:0.019446396043273483, epicenter:0.019033877626552984, kashmir:0.017538077834802355, younghusband:0.01662707726890508
#1235	Chile:0.9838234058958835, earthquake:0.7334719238518906, hubo:0.18330428936195908, depth:0.17423672076764093, offshore:0.1531609591913485, 11:0.12586171637712218, pais:0.12165389943213505, concepcion:0.09873758928190768, egypt:0.09451777487705819, unico:0.09404993567479851, xd:0.09120271136821441, feb:0.08603458827462157, afar:0.08130600624167321, dice:0.07495004576927244, eso:0.0744289849636961, ago:0.07037190450699154, bio:0.06898210978764117, told:0.06563079193341446, nw:0.06556320156610507, alguien:0.061061490096581815
#3696	Japan:37.46591519678831, earthquake:27.79528941386803, hawaii:13.52850900901365, warning:9.81412145274017, tokyo:8.204033395331049, quake:7.451195147758577, plant:7.325899561790927, prayers:6.9026673646340875, affected:6.584122684154533, philippines:6.0432670900161405, philippines:6.0432670900161405, jepang:5.6354616652034295, victims:5.042144511805812, coast:4.508437995331032, pray:4.017769359000981, cnn:3.8654987635502525, issued:3.7513719220338917, indonesia:2.9603077154240593, praying:2.7032789033476887, japanese:2.629934771164886, relief:2.442524279343379
#226	Haiti:0.42992990944228393, earthquake:0.2409183229418004, video:0.0989920511133773, money:0.08840308705031176, mobile:0.08071766464658471, anniversary:0.07855424457000261, year:0.07226845690067119, hope:0.07017435839712309, loads:0.06362512372035956, introduction:0.061828265596069264, ide:0.04130928012277363, out:0.035292134498553, check:0.025520818129676263, lost:0.013946163044925304, smh:0.011884986501718282, facebook:0.011481708528178978, gut:0.005129003776993684, remember:0.0049169377155100285, 49:0.0040700118820341455, app:0.0037549258587891553
#5994	Magnitude:0.6629561016669671, spain:0.4853259248622711, earthquake:0.3207357341247239, spanish:0.1317980996902689, killed:0.12636632831607628, southern:0.12398514886803128, eastern:0.11867037958047656, town:0.09087563528134832, lorca:0.087688354129441 56, hit:0.07739086244497273, 10:0.07150847488846834, wednesday:0.06898969023614157, people:0.06863952761038158, dead:0.0662505262577082, rocks:0.057081708958353225, hits:0.05222377115382082, south:0.0493579328504147, reuters:0.049191006364638676, deadly:0.048013608253639616, video:0.04551498183728966
#3983	Chile:0.28500723889101887, depth:0.1530729350043406, 16:0.13731867977513712, earthquake:0.13436252264397855, mar:0.10886923475005959, 25km:0.09230892510857894, 36pm:0.08614837251790954, ago:0.07480396352065272, epicenter:0.04066824479683694, nne:0.04053257546213937, valparaiso:0.017291041090262204, 72km:5.653316665865345E-4, 14m:5.653316665865345E-4
#4329	Thailand:0.8069340289763622, magnitude:0.42674047743767, earthquake:0.2567410486251388, rai:0.25038277300578704, north:0.16993467719435945, survey:0.14889847036249787, reuters:0.12866723872570304, 69:0.11396669287525248, extensive:0.09162587045384778, miles:0.08905338877167877, allah:0.083429068587796327, chiang:0.08256460369605129, hope:0.06951695314927987, myanmar:0.06005008186458635, hits:0.059245532262360324, brlaku:0.059004617752693694, geological:0.0548152346821232, earthquakes:0.039566581181361325, dgn:0.03437806676884497, td:0.03175587332167133
#5647	Obama:12.745330253889234, dead:7.628842394409808, bush:5.727640210614691, cnn:5.512015097438911, muerte:5.403916698731788, killed:5.365761326438391, troops:4.55536125496666, president:4.105680930699426, seek:3.9013848774228705, hide:2.9769707486473123, confirmed:2.885231345097644, speech:2.5849019550065986, pakistan:2.4023648359349052, america:2.3983367065905217, reporting:2.3957262704792006, announce:2.3243426866964896, justice:2.290096565848105, hitler:2.2778369257657136, announcement:2.186014823436279, death:2.1546555923826047
#1339	Eminem:11.136926274122597, usher:7.424567248673083, gaga:6.703801172105693, rihanna:5.9146758836688615, grammys:5.256234271217888, drake:4.6893691868933365, katy:4.513606054965342, dre:4.419741601924677, christina:4.302273677133062, egg:4.029607088962011, bruno:3.835773522512244, valentines:3.803607350486565, performance:3.7615012201679217, jennifer:3.711795303426252, smith:3.591422768289237, carpet:3.4254347795835085, grammy:3.3937668242901675, perry:3.293253227668383, lady:3.0041180756508776, valentine:2.969949127529637
#1021	Peas:11.959902209325707, usher:9.74856012495839, eyed:8.343619330107162, christina:8.32598979550528, packers:6.770466602414572, superbowl:5.96686744475942, steelers:5.263423502122333, commercial:4.607126071523586, bowl:4.05503816366219, eminem:3.9407917852968124, glee:3.750948911288515, yellow:3.7170796188059643, bay:3.5909152562193402, commercials:2.552539269333649, halftime:2.2987344388722013, aaron:2.197443500298245, green:2.150154047796244, tron:1.5792979751751852, detroit:1.4931286316980692, performance:1.4770572366430026
#5853	Lakers:11.042592853678425, mothers:8.814775357888216, phil:4.042779209644956, kobe:2.97709501940586, terry:2.585166153770605, chelsea:2.3462281832106613, moms:2.0462373302632404, dallas:1.7936280114467864, mavs:1.4670466391613874, mom:1.169273799087787, jackson:1.0716962221156736, mother:0.7899717293201031, mommy:0.7593037231261597, happymothersday:0.7355215975788701, bench:0.7167117631345038, glory:0.7018865014144192, sweep:0.6912593603295222, happy:0.6856431911484824, jason:0.5919595667468056, 86:0.5754961457163142
#3115	Inception:11.341717888099115, oscar:10.222144566847703, portman:10.094859784347854, natalie:9.881543882505577, anne:6.186457828186518, oscars:5.010312361328005, carpet:3.4736367060626936, alice:3.1757714768745267, christian:2.8546228150439563, oprah:2.8416152456987347, toy:2.7703290002097822, actor:2.289160612701109, james:1.923834962856439, teresa:1.881646263732395, billy:1.7424228883730524, crystal:1.681513187170654, speech:1.6380498467290179, bale:1.601160642161866, nicole:1.560718452491151, fighter:1.5315258062366097
#5536	Kate:31.860395069795477, william:22.910384439151454, royal:15.145100058085736, wedding:11.48109099089283, prince:11.34096005074171, royalwedding:6.6272453508469304, harry:5.145489509136396, queen:5.117233791763261, diana:4.183708813148399, draft:3.5641875805264185, dress:3.46606892121761, middleton:3.2365724924230834, princess:3.000560653182693, saints:2.5431067464730828, kiss:2.2786574214066233, hats:2.1673340721319336, alexander:2.132957432840678, stunning:2.1245806306756205, palace:2.0585690024252585, nikah:1.8499794949849484

of event was detected by our system at 02:02:17 and 10:57:13 is illustrated in Table 2. The resulting related events detected by our system (per 10 min) are illustrated in Fig. 7. The top 10 keywords and their burst weighting of events is illustrated in Table 1.

- **Results and discussion** (Ranking of related events).

We utilized Virginia earthquake event (August 24, 2011 and original event ID is #1173) as our baseline to testify our framework. In our system, the Virginia earthquake was detected at 01:52:10 on August 24. This event was compared with the collection of formulated events per ten minute. In Table 2, the related event compared with baseline event is illustrated. The top one of resulting events is Christchurch earthquake. This is perhaps because that these two events in common both earthquakes occurred in city areas and both had aftershocks.

6.1.2. Case study (II): baseline event "Whitney Houston dead"

In the experiment, a total number of 575,438,311 Twitter posts were collected, dating from: October 1, 2010 to September 30, 2011 and January 1, 2012 to March 14, 2012. After filtering out non-ASCII tweets, 304,758,200 tweets had been utilized as our data source. Also, we utilized the dataset collected from October 1, 2010 to September 30, 2011 corpus as our dataset for training, and used the corpus dating from January 1, 2012 to March 14, 2012 as our test data. Subsequently, we partitioned messages into unigrams and all capital letters in each tweet were converted into lowercase for our experiments.

- **The datasets** (Case II: baseline event "Whitney Houston dead on February 12, 2012").

In this case, a total number of 116,081 Twitter posts associated with "Whitney Houston dead" event (Event ID: #1318, February 12, 2012) were collected, dating from: February 12, 2012 to February 13, 2012. The first message related to the event was posted on February 12 08:57:38 CST 2012. Such an event was detected by our system on February 12 08:58:51 CST 2012.

- **Results and discussion** (Ranking of related events).

We utilized Whitney Houston dead event (February 12, 2012 and original event ID is #1318) as our baseline to testify our framework. In our system, the event was detected at 08:58:51 on February 12. The result of relatedness ranking of event was computed by our system at February 12 09:18:50 and February 13 16:16:40 respectively, as illustrated in Table 4. The resulting related events detected by our system (per 10 min) are illustrated in Fig. 8. The top 10 keywords and their burst weighting of events is illustrated in Table 3.

This event was compared with other detected events per ten minute. In Table 4, compared with baseline event the most related event is *Amy Winehouse dead* event at February 12 09:18:50. After that, the most related event became *MTV video music awards* event at February 13 16:16:40. This is perhaps because the event is more related to the fields of entrainment such as singers and music activity.

6.2. Experimenting with offline event evaluation

To enable the investigation of large-data solutions to event-relatedness modeling, we have collected a total number of 575,438,311 Twitter posts, dating from October 1, 2010 to September 30, 2011 and January 1, 2012 to March 14, 2012. The test samples were collected through Twitter Stream API. After filtering out non-ASCII tweets, 304,758,200 tweets had been utilized as our data source. We utilized the dataset collected from October 1, 2010 to September 30, 2011 as our corpus for training, and used the

Table 2
Illustration of ranking of related events upon a comparison with a baseline. "Virginia earthquake" event at various time points (Event ID: #1173, August 24, 2011).

Relatedness (%)	Event	Human judgment
Ranking event at 02:02:17 on August 24, 2011		
73.615	Event ID: #2309, Christchurch Earthquake (February 22, 2011)	**0.9**
73.219	Event ID: #398, Pakistan Earthquake (January 19, 2011)	0.84
60.748	Event ID: #4204, Philippines Earthquake (March 21, 2011)	0.76
55.797	Event ID: #1235, Chile Earthquake (February 12, 2011)	0.8
50.199	Event ID: #3696, Japan Earthquake (March 11, 2011)	0.5
44.199	Event ID: #226, Haiti Earthquake (January 13, 2011)	0.5
33.632	Event ID: #5994, Spain Earthquake (March 12, 2011)	0.8
31.704	Event ID: #3983, Chile Earthquake (March 17, 2011)	0.74
24.424	Event ID: #4329, Thailand Earthquake (March 24, 2011)	0.8
0.146	Event ID: #1339, Grammy Award (February 14, 2011)	0.1
0.0571	Event ID: #1021, Superbowl (February 06, 2011)	0.14
0.04	Event ID: #5647, Osama Bin Laden Dead (May 02, 2011)	0.18
0.03	Event ID: #5853, Happy Mother's Day (May 08, 2011)	0.2
0.03	Event ID: #3115, Oscar (February 28, 2011)	0.16
0.008	Event ID: #5536, Royal Wedding (April 28, 2011)	0.12
Ranking event at 10:57:13 on August 24, 2011		
71.665	Event ID: #2309, Christchurch Earthquake (February 22, 2011)	0.9
71.139	Event ID: #398, Pakistan Earthquake (January 19, 2011)	0.84
60.127	Event ID: #4204, Philippines Earthquake (March 21, 2011)	0.76
54.23	Event ID: #1235, Chile Earthquake (February 12, 2011)	0.8
50.116	Event ID: #3696, Japan Earthquake (March 11, 2011)	0.5
42.899	Event ID: #226, Haiti Earthquake (January 13, 2011)	0.5
32.97	Event ID: #5994, Spain Earthquake (March 12, 2011)	0.8
30.737	Event ID: #3983, Chile Earthquake (March 17, 2011)	0.74
24.106	Event ID: #4329, Thailand Earthquake (March 24, 2011)	0.8
1.248	Event ID: #5647, Osama Bin Laden Dead (May 02, 2011)	0.18
0.676	Event ID: #1339, Grammy Award (February 14, 2011)	0.1
0.342	Event ID: #1021, Superbowl (February 06, 2011)	0.14
0.282	Event ID: #5853, Happy Mother's Day (May 08, 2011)	0.2
0.243	Event ID: #3115, Oscar (February 28, 2011)	0.16
0.169	Event ID: #5536, Royal Wedding (April 28, 2011)	0.12

collected messages dating from January 1, 2012 to March 14, 2012 as our test data. Most traditional data categorization systems use a single categorization procedure to determine the topic category of a given dataset. However, for event-related messages involving a number of topics and noisy inputs, it is difficult to employ the above way to differentiate their relatedness for analysis. This has led to a sensible solution through comparing the social-media contents associated with specific events in the way of several trained topic-classifiers to support discriminating the extent of involvement in each predefined topic category for gauging the impacts of the events. As stated previously, through formulating the collection of related social messages an event-story can be modeled by a number of hidden topics and selected features, with each topic gathering a series of observed messages according to topic-specific terms and sentences used in the content. A representation of the event comparison can be done by the quantitative values represented in each defined topic-dimension in a classifier (topic)-based space in a supervised manner. This measures the model's descriptive power, while requiring no submitted on-going event data for evaluation. As a result, in this work we attempt to explore the capacity of a multiple classifier system in dealing with the evaluation of event relatedness.

6.2.1. The learned model of topic classifiers

In order to establish an offline learned model to quantitatively measure the relatedness of event impacts, we study on analyzing the contents of Twitter datasets related to the events. In this work, the experimental process includes two phases. First, we are focusing on the generation of eight well-trained SVM classifiers by means of training with the Twitter messages of selected topics, covering *Politics*, *Business*, *Finance*, *Climate*, *Commodity*, *Health*, *Entertainment*, and *Sport* topic domains. To collect the corpus for training, we crawled Twitter messages using its publicly available API. For each SVM classifier, we utilized millions of topic specific messages for training, and numerous messages for testing operations. The topic classifiers were well developed based upon the best results performed by the training and testing process mentioned above. Subsequently, we started to perform relatedness

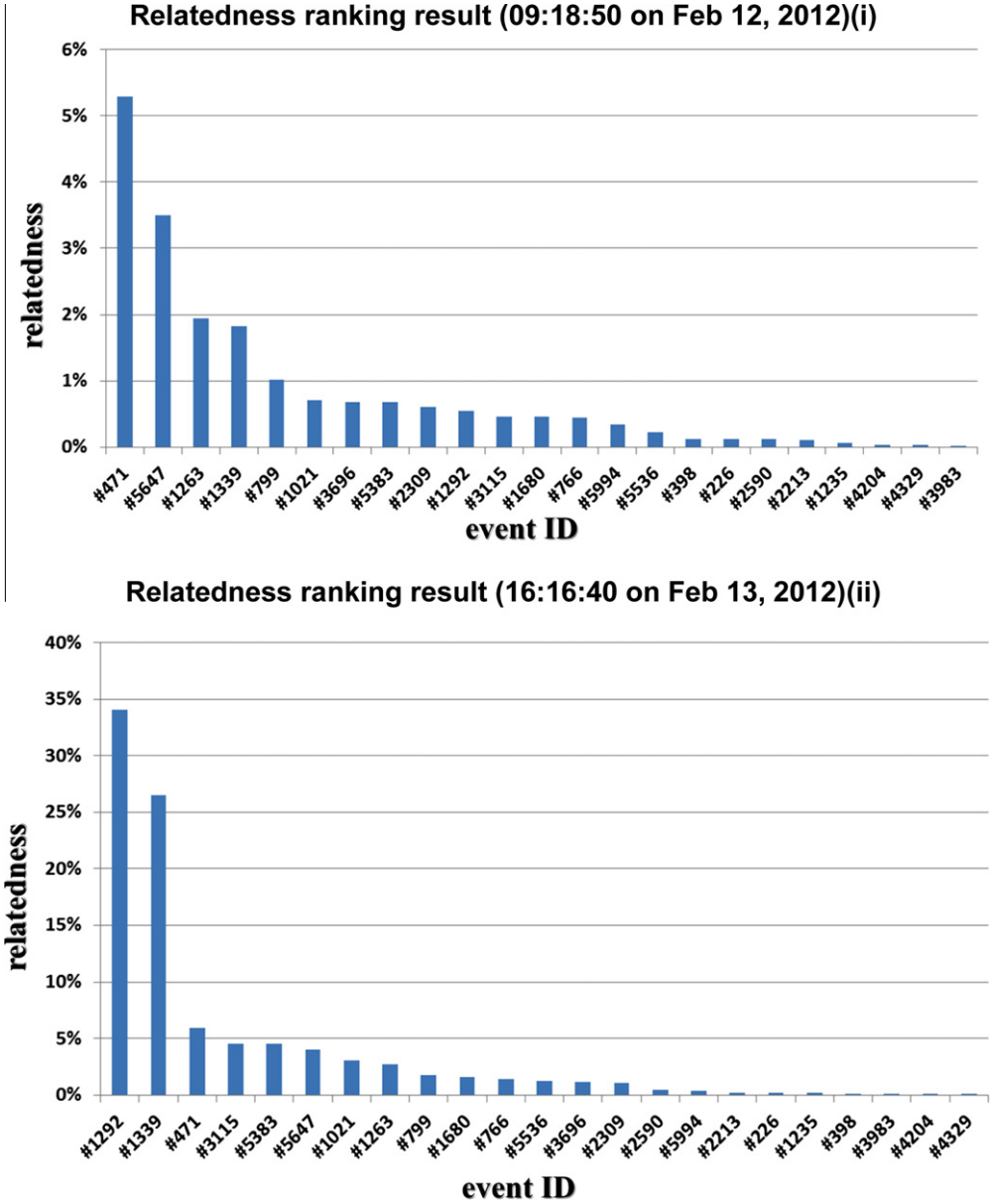


Fig. 8. Ranking of event relatedness upon a comparison with baseline "Whitney Houston dead" event at various time points (Event ID: #1318, February 12, 2012).

Table 3
Top 10 keywords and the burst weighting in the detected events.

Event ID	Keywords and burst weighting
#1318	Whitney:18.43391694588743, houston:13.00238271313586, adele:9.118528622879799, nicki:7.852207140486721, hudson:6.467371490829314, minaj:5.445753126578573, jennifer:5.2844432315326655, alicia:4.454457380222075, rihanna:4.349160148846695, swift 4.262224975323266,
#2309	Earthquake:3.9553197961748126, zealand:0.963100917277774, prayers:0.909794403440777, 0800:0.884466396150691, affected:0.8383470931121912, nz:0.8314472900536037, cricket:0.7336453038423575, 65:0.6659161651840394, 22:0.6243486226423245, coverage:0.6216531727745904
#398	Earthquake:1.04908551454045, pakistan:0.8980141224744113, hits:0.16401090117859768, 18:0.15131980804529255, sw:0.12150740246713974, western:0.10616833459240625, major:0.09292551543064907, south:0.09152100337498355, means:0.08779825561221398, hit:0.08275405837041092
#4204	Philippines:0.49761490889948634, earthquake:0.47233581624246196, azkals:0.24853693404998034, safe:0.10928621389049406, manila:0.09546108516851091, birthday:0.08997353333089236, felt:0.08747401235101598, wnw:0.08061196314147687, 21:0.07613447410948651, magnitude:0.07176278026603079
#1235	Chile:0.9838234058958835, earthquake:0.7334719238518906, hubo:0.18330428936195908, depth:0.17423672076764093, offshore:0.1531609591913485, 11:0.12586171637712218, pais:0.12165389943213505, concepcion:0.09873758928190768, egypt:0.09451777487705819, unico:0.09404993567479851
#3696	Japan:37.46591519678831, earthquake:27.79528941386803, hawaii:13.52850900901365, warning:9.814121452740 17, tokyo:8.204033395331049, quake:7.451195147758577, plant:7.325899561790927, prayers:6.9026673646340875, affected:6.584122684154533, philippines:6.0432670900161405
#226	Haiti:0.42992990944228393, earthquake:0.2409183229418004, video:0.0989920511133773, money:0.08840308705031176, mobile:0.08071766464658471, anniversary:0.07855424457000261, year:0.07226845690067119, hope:0.07017435839712309, loads:0.06362512372035956, introduction:0.061828265596069264
#5994	Magnitude:0.6629561016669671, spain:0.4853259248622711, earthquake:0.3207357341247239, spanish:0.1317980996902689, killed:0.12636632831607628, southern:0.12398514886803128, eastern:0.11867037958047656, town:0.09087563528134832, lorca:0.08768835412944156, hit:0.07739086244497273
#3983	Chile:0.28500723889101887, depth:0.1530729350043406, 16:0.13731867977513712, earthquake:0.13436252264397855, mar:0.10886923475005959, 25km:0.09230892510857894, 36pm:0.08614837251790954, ago:0.07480396352065272, epicenter:0.04066824479683694, nne:0.04053257546213937
#4329	Thailand:0.8069340289763622, magnitude:0.42674047743767, earthquake:0.2567410486251388, rai:0.25038277300578704, north:0.16993467719435945, survey:0.14889847036249787, reuters:0.12866723872570304, 69:0.11396669287525248, extensive:0.09162587045384778, miles:0.08905338877167877
#5647	Obama:12.745330253889234, dead:7.628842394409808, bush:5.727640210614691, cnn:5.512015097438911, muerte:5.403916698731788, killed:5.365761326438391, troops:4.55536125496666, president:4.105680930699426, seek:3.9013848774228705, hide:2.9769707486473123
#1339	Eminem:11.136926274122597, usher:7.424567248673083, gaga:6.703801172105693, rihanna:5.9146758836688615, grammys:5.256234271217888, drake:4.6893691868933365, katy:4.513606054965342, dre:4.419741601924677, christina:4.302273677133062, egg:4.029607088962011
#1021	Peas:11.959902209325707, usher:9.74856012495839, eyed:8.343619330107162, christina:8.32598979550528, packers:6.770466602414572, superbowl:5.96686744475942, steelers:5.263423502122333, commercial:4.607126071523586, bowl:4.05503816366219, eminem:3.9407917852968124
#3115	Inception:11.341717888099115, oscar:10.222144566847703, portman:10.094859784347854, natalie:9.881543882505577, anne:6.186457828186518, oscars:5.010312361328005, carpet:3.4736367060626936, alice:3.1757714768745267, christian:2.8546228150439563, oprah:2.8416152456987347
#5536	Kate:31.860395069795477, william:22.910384439151454, royal:15.145100058085736, wedding:11.48109099089283, prince:11.34096005074171, royalwedding:6.6272453508469304, harry:5.145489509136396, queen:5.117233791763261, diana:4.183708813148399, draft:3.5641875805264185
#1292	Vma:10.935486385400269, adele:9.746942707153927, jessie:7.936963732817145, nicki:7.141767744487609, gaga:5.923693949264141, minaj:5.449700397326526, britney:4.432865885476575, katy:4.312670805530427, beyonce:3.9796277208499813, kanye:3.834185543454056
#471	Amy:18.801063607730192, rip:5.931832341943809, rehab:3.8761460430726085, singer:2.4481710285827147, 27:2.2211722783856533, sources:1.9019016212417, talented:1.7893147230208886, died:1.7383830689271476, dead:1.642763226317653, azkals:1.3625730473296502
#799	Birmingham:2.1162404051429 79, cameron:1.7844374654616597, london:1.5692771988530108, police:0.9923891434255501, army:0.9789689381375396, liverpool:0.6675253938471393, rioters:0.6665477287429897, safe:0.6295441758244948, shops:0.5525338051098853, burning:0.5373842519398319
#1263	Hurricane:1.639764934872068, liverpool:0.8668768518891237, storm:0.5097137350428569, irene:0.4283028053395561, respects:0.42217586177080424, bolton:0.41020409698649785, ynwa:0.40376746590619, lfc:0.3777959373225377, power:0.33906504876268534, tornado:0.31451102137507625
#5383	Easter:5.909993992493764, church:2.517663462702729, eggs:2.0321252962831413, happy:1.1073846666972311, pascoa:1.0231833433907362, bunny:1.0163247884192292, risen:1.0106248612150663, egg:0.9892130157054428, celtics:0.8459297222404398, jesus:0.7420455376658811
#2590	Bola:6.12194766185904, indonesia:5.828293782224222, menang:4.4334391637808, timnas:4.223264557991573, pemain:3.722917263773142, markus:3.005078027340159, gol:2.9820639777224915, irfan:2.1321614674161933, ganteng:2.0094005682788927, bepe:1.8792127828299796
#2213	Thailand:24.606029550383553, bp:6.872369041311373, indonesia:6.321821319068359, menang:5.5163776567442415, thai:4.618014365446932, thai:4.618014365446932, gol:3.5563785024284673, ina:1.9592020192138961, kalah:1.956115919396642, pemain:1.5389320733667542
#766	Dilma:5.51226615657667, saints:3.7455014738070944, candy:3.450429149343459, trick:3.3491944256538106, halloween:3.2790728991827587, steelers:2.9681353056313773, presidente:1.8596312239437545, aiden:1.5752707381107498, brasil:1.5722256034663376, jets:1.184611555030256
#1680	Thanksgiving:13.28027687567116, thankful:10.62765828432697, beyonce:7.2430940196459, turkey:5.446051031150933, saints:4.4011402428229704, lions:3.99631348997024, parade:3.3759528361915256, cowboys:2.6267613585885052, happy:2.171525486003451, walmart:1.5495834866010616

evaluation on the platform of the multiple classifiers, and the event datasets were being formulated and go through the identification procedure of categorical decision process of the classifiers to produce formulated event-vectors. The resulting vectors represent the essential features of the respective events, based on the judgments of employed multiple classifiers. Finally, we measured the relatedness of event vectors through the Distance, Cosine, Dice and Jacard measuring methods (Salton, 1989). The examples of experimental results are presented in the following sections.

6.2.2. Experimental results

In this section the examples of experimental results are demonstrated. According to the resulting decision ratings of respective event vectors, we can then perform several measures of relatedness between event pairs, by means of Cosine, Distance, Dice and Jacard methods. Thus, the degree of relatedness between tested event vectors can be obtained, as shown in Figs. 9–13.

6.2.3. Evaluation

To testify the experimental results, we compare the results of measures with human judgments. Although there is no standard way to evaluate computational measures of text relatedness, one sensible way to judge would seem to be agreement with human relatedness ratings. This can be assessed by means of a computational relatedness measure to rate the relatedness of a set of document pairs, and looked at how well its ratings correlate with human ratings of the same pairs. In our evaluation process, 20 human reviewers were asked to rate "relatedness of text contents" for

Table 4
Relatedness ranking of event based upon baseline event "Whitney Houston dead" (Event ID: #1318, February 12, 2012).

Relatedness (%)	Event	Human judgment (%)
Ranking event at 09:18:50 on Feb 12, 2012(i)		
5.2971	Event ID: #471, Amy Winehouse dead (Jul 23, 2011)	**50**
3.5003	Event ID: #5647, Obama Says Osama Bin Laden Is Dead (May 02, 2011)	20
1.9404	Event ID: #1263, Hurricane Irene (Aug 27, 2011)	10
1.8190	Event ID: #1339, Grammy Awards (Feb 14, 2011)	10
1.0122	Event ID: #799, London Rioting (Aug 09, 2011)	10
0.7106	Event ID: #1021, Superbowl (Feb 06, 2011)	10
0.6847	Event ID: #3696, Japan Earthquake (Mar 11, 2011)	10
0.6734	Event ID: #5383, Happy Easter (Apr 23, 2011)	10
0.6134	Event ID: #2309, New Zealand Earthquake (Feb 22, 2011)	10
0.5508	Event ID: #1292, MTV Video Music Awards (Aug 28, 2011)	10
0.4607	Event ID: #3115, Oscar (Feb 28 , 2011)	10
0.4566	Event ID: #1680, Thanksgiving Day (Nov 24, 2011)	10
0.4428	Event ID: #766, Halloween (Oct 30, 2011)	10
0.3492	Event ID: #5994, Spain Earthquake (May 12, 2011)	10
0.2264	Event ID: #5536, Royal Wedding (Apr 28, 2011)	10
0.1186	Event ID: #398, Pakistan Earthquake (Jan 19, 2011)	10
0.1164	Event ID: #226, Haiti Earthquake (Jan 13, 2011)	10
0.1163	Event ID: #2590, Philippine vs Indonesia (Dec 16, 2010)	10
0.1091	Event ID: #2213, Indonesia vs Thailand (Dec 07, 2010)	10
0.0602	Event ID: #1235, Chile Earthquake (Feb 12)	10
0.0360	Event ID: #4204, Philippines Earthquake (Mar 21, 2011)	10
0.0291	Event ID: #4329, Thailand Earthquake (Mar 24, 2011)	10
0.0181	Event ID: #3983, Chile Earthquake (Mar 17, 2011)	10
Ranking event at 16:16:40 on February 13, 2012(ii)		
34.0740	Event ID: #1292, MTV Video Music Awards (Aug 28, 2011)	30
26.4876	Event ID: #1339, Grammy Awards (Feb 14, 2011)	**38**
5.9119	Event ID: #471, Amy Winehouse dead (Jul 23, 2011)	18
4.5193	Event ID: #3115, Oscar (Feb 28 , 2011)	16
4.5193	Event ID: #5383, Happy Easter (Apr 23, 2011)	10
4.0217	Event ID: #5647, Obama Says Osama Bin Laden Is Dead (May 02, 2011)	10
3.0910	Event ID: #1021, Superbowl (Feb 06, 2011)	10
2.7557	Event ID: #1263, Hurricane Irene (Aug 27, 2011)	10
1.7480	Event ID: #799, London Rioting (Aug 09, 2011)	10
1.5773	Event ID: #1680, Thanksgiving Day (Nov 24, 2011)	10
1.4211	Event ID: #766, Halloween (Oct 30, 2011)	10
1.2643	Event ID: #5536, Royal Wedding (Apr 28, 2011)	10
1.1687	Event ID: #3696, Japan Earthquake (Mar 11, 2011)	10
1.0943	Event ID: #2309, New Zealand Earthquake (Feb 22, 2011)	10
0.4344	Event ID: #2590, Philippine VS Indonesia (Dec 16, 2010)	10
0.3293	Event ID: #5994, Spain Earthquake (May 12, 2011)	10
0.2299	Event ID: #2213, Indonesia VS Thailand (Dec 07, 2010)	10
0.1675	Event ID: #226, Haiti Earthquake (Jan 13, 2011)	10
0.1600	Event ID: #1235, Chile Earthquake (Feb 12)	10
0.1442	Event ID: #398, Pakistan Earthquake (Jan 19, 2011)	10
0.1348	Event ID: #3983, Chile Earthquake (Mar 17, 2011)	10
0.1164	Event ID: #4204, Philippines Earthquake (Mar 21, 2011)	10
0.0753	Event ID: #4329, Thailand Earthquake (Mar 24, 2011)	10

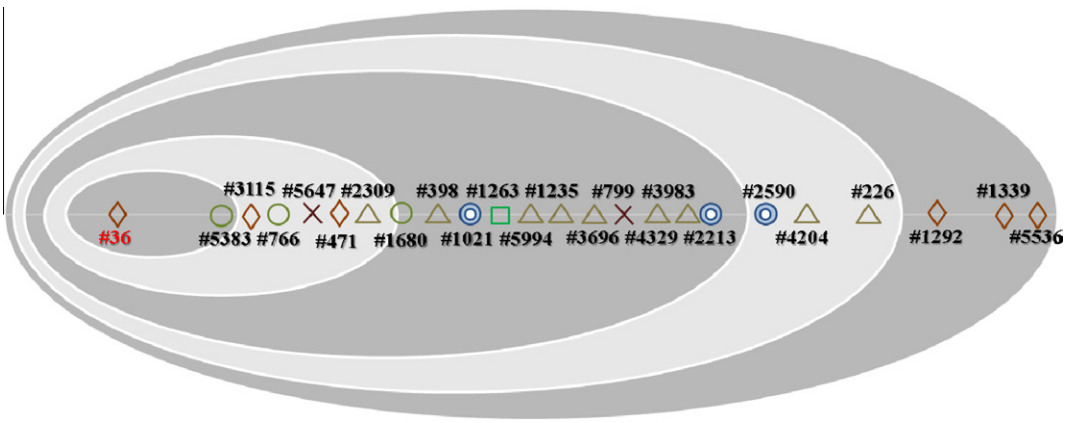


Fig. 9. Representation of an example measure of event relatedness (1).

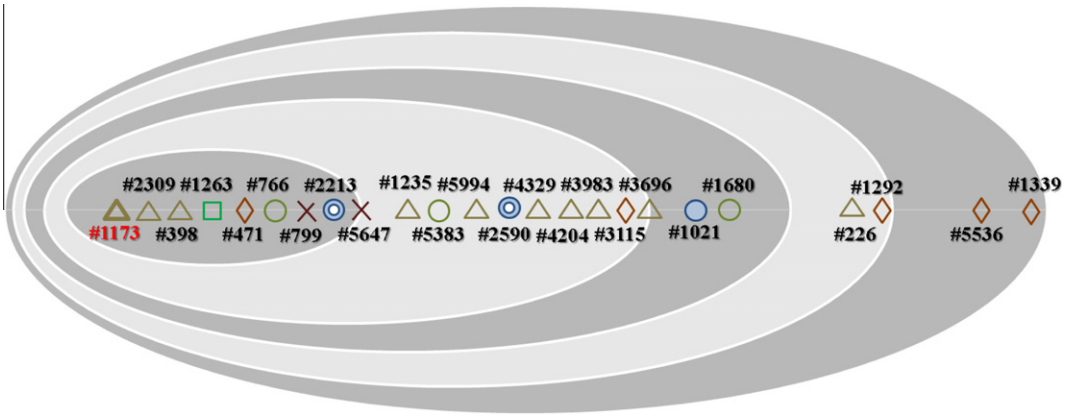


Fig. 10. Representation of an example measure of event relatedness (2).

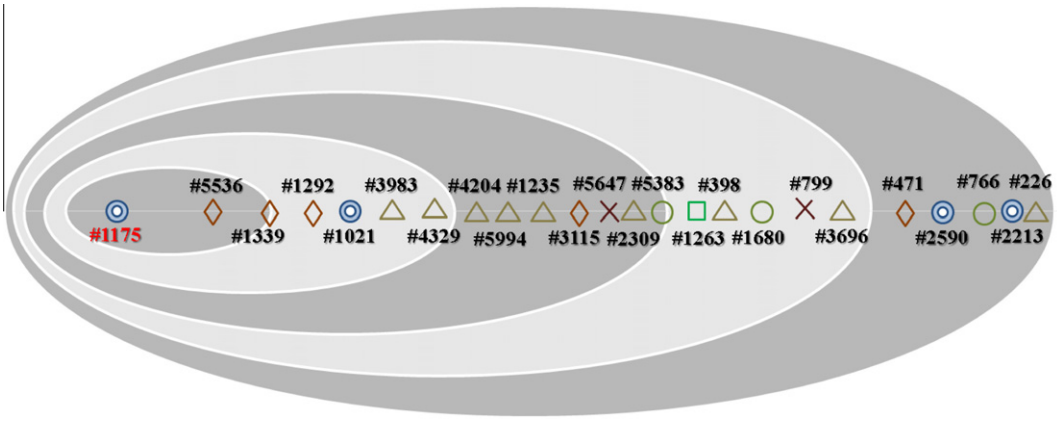


Fig. 11. Representation of an example measure of event relatedness (3).

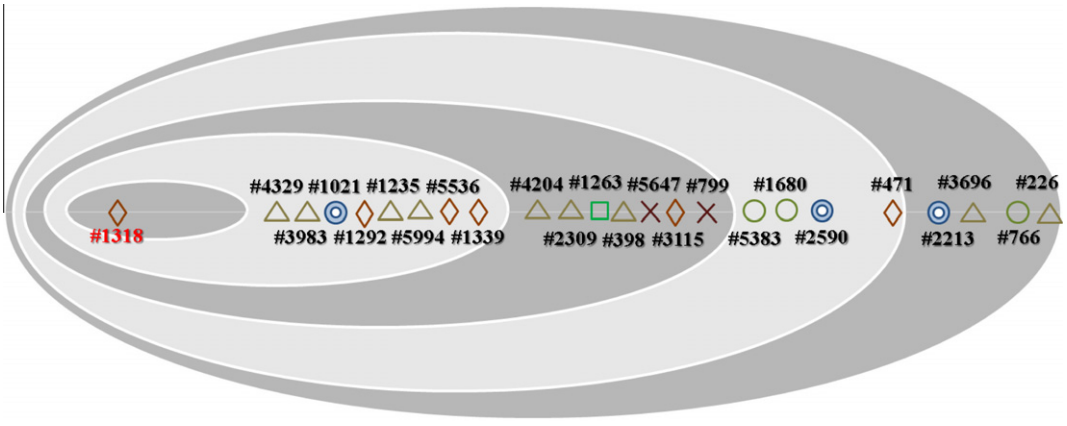


Fig. 12. Representation of an example measure of event relatedness (4).

each pair on a scale from 0 (no relatedness) to 100% (perfect). The average rating for each pair thus represent a good estimate of how related the two documents are, according to human judgments. Table 7 summarizes the experimental results, giving the correlation between the computational relatedness ratings and the mean ratings by measures of human judgments. In Table 7, some of example results of the measures of event relatedness including Distance, Cosine, Dice and Jacard, appear to be agreement with human relatedness ratings. This suggests that to some extent the model using the multi-classifier method provides results that are sensible and useful for event relatedness measures and analysis.

Once all tested events have been computed the extent of impacts on several topic domains on our multiple classifier system, we started to identify the effects of timing factor and geographic location on the experimented events. To indicate the timing the events, in Table 6 we used Boolean value (i.e., '0' or '1') to represent the event occurring time represented by quarter (i.e. Q4, Q1, Q2, Q3). In Table 6, Q4 represents occurring time dating from October

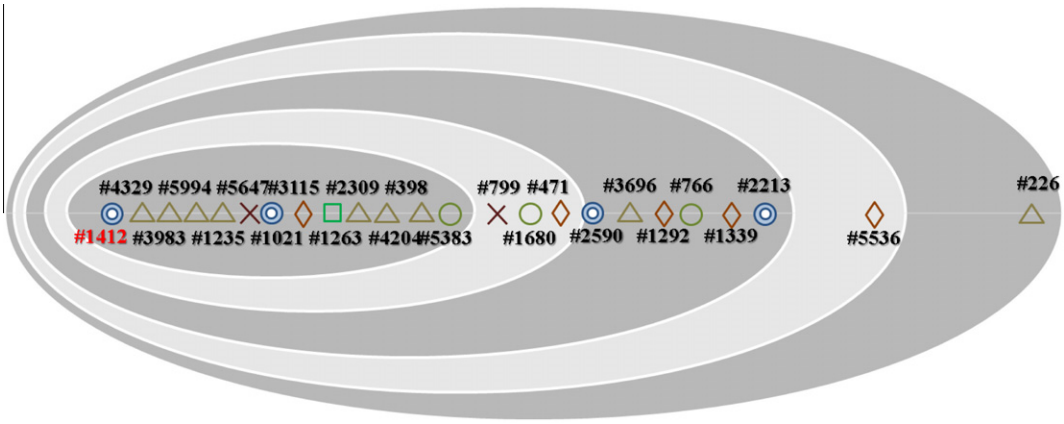


Fig. 13. Representation of an example measure of event relatedness (5).

Table 5
Top 10 countries (measured by time zone) of the volume of experimental messages.

Country	Number of messages	Percentage
USA	3,161,310	52.6885
Ecuador	532,022	8.86703
Chile	492,410	8.20683
Brazil	447,230	7.45383
Indonesian	425,948	7.09913
UK	296,310	4.9385
Netherlands	186,706	3.11176
Singapore	75,322	1.25536
Australia	48,094	0.8015
Japan	34,730	0.57883

01, 2010 to December 31, 2010, Q1 represents occurring time dating from January 01, 2011 to March 31, 2011 (#3115, #1021, #2309, #398, #4204, #1235, #3696, #226, #5994, #3983, #4329, #1339) and dating from January 01, 2012 to March 31, 2012 (#36, #1175, #1318, #1412), Q2 represents occurring time dating from April 01, 2011 to June 30, 2011, and Q3 represents occurring time dating from July 01, 2011 to September 30, 2011. The values listed below the field of countries were generated by calculating the proportion of location (i.e. time zone) of event-messages. When looking at the effect of geographic location we use time zone represented by national boundaries (i.e. countries) to indicate the locations of the events. It is clear that not all nations are created equal. Initially, the events might not occur in the United States,

Table 6
The spatio-temporal information of events.

Event ID	Spatio-temporal feature													
	Time				Geographic location (time zone)									
	Q4 (10, 11, 12)	Q1 (1, 2, 3)	Q2 (4, 5, 6)	Q3 (7, 8, 9)	Japan	Australia	Singapore	Netherlands	UK	Indonesian	Brazil	Chile	Ecuador	USA
#1263	0	0	0	1	0.001999	0.001499	0.002499	0.003248376	0.013993	0.006247	0.003498	0.003498	0.185157	**0.535482**
#799	0	0	0	1	0.00181	0.010179	0.001131	0.006559602	0.376838	0.00656	0.003619	0.00475	0.024881	0.137073
#471	0	0	0	1	0.003463	0.002837	0.006008	0.022905541	0.07026	0.012392	0.074307	0.071303	0.062876	**0.285255**
#1292	0	0	0	1	0.001256	0.00162	0.003134	0.007947595	0.025322	0.005635	0.050303	0.059425	0.136682	**0.418499**
#3115	0	1	0	0	0.001308	0.006214	0.002756	0.002336012	0.019062	0.012287	0.077135	0.043263	0.058027	**0.47122**
#5383	0	0	1	0	0.002743	0.006171	0.004656	0.004078097	0.053484	0.027428	0.0179	0.017828	0.077845	**0.465589**
#5647	0	0	1	0	0.003144	0.007508	0.003998	0.002594311	0.012453	0.006471	0.021487	0.014253	0.092632	**0.590984**
#1021	0	1	0	0	0.001404	0.002439	0.000912	0.002315727	0.023404	0.00239	0.015028	0.010519	0.091816	**0.607213**
#5536	0	0	1	0	0.00412	0.01962	0.015601	0.01105361	0.135557	0.064689	0.009019	0.005879	0.037708	**0.338919**
#2590	1	0	0	0	0.003025	0.001945	0.005618	0.000432152	0.002161	0.351556	0.000216	0.000432	0.001729	**0.435825**
#2213	1	0	0	0	0.001665	0.001295	0.005736	0.001110289	0.002591	0.342709	0.000185	0.000555	0.00074	**0.446151**
#766	1	0	0	0	0.003516	0.002792	0.004447	0.005584859	0.038577	0.009825	0.10001	0.06464	0.06857	**0.419071**
#1680	1	0	0	0	0.002743	0.002743	0.002939	0.003056426	0.015674	0.004702	0.004624	0.003801	0.103605	**0.587422**
#2309	0	1	0	0	0.003597	0.179856	0.028777	0	0.093525	0.007194	0.003597	0.017986	0.010791	**0.338129**
#398	0	1	0	0	0	0	0	0	0.101449	0	0	0.028986	0	**0.42029**
#4204	0	1	0	0	0	0	0.148148	0	0	0	0	0	0	**0.444444**
#1235	0	1	0	0	0	0	0	0	0	0	0	0.5	0.05	0.125
#3696	0	1	0	0	0.038666	0.011819	0.020583	0.016870391	0.055662	0.059804	0.007551	0.009521	0.036418	**0.413855**
#226	0	1	0	0	0	0.015385	0	0	0.030769	0.015385	0	0.015385	0.061538	**0.6**
#5994	0	1	0	0	0.015873	0.015873	0	0.031746032	0.15873	0.015873	0	0	0	**0.31746**
#3983	0	1	0	0	0	0	0	0	0	0	0	0.777778	0	0.055556
#4329	0	1	0	0	0.066667	0	0.083333	0	0.05	0.066667	0	0.016667	0.033333	**0.25**
#1339	0	1	0	0	0.001847	0.003767	0.004115	0.00341971	0.02968	0.014319	0.029388	0.025675	0.100452	**0.516266**
#1173	0	0	0	1	0.002094	0.000739	0.000739	0.001909337	0.012257	0.001293	0.002279	0.003141	0.187669	**0.597561**
#36	0	1	0	0	0.001309	0.005832	0.003095	0.002856463	0.018924	0.004166	0.020233	0.024637	0.091169	**0.507617**
#1175	0	1	0	0	0.000662	0.00205	0.000299	0.00109964	0.019185	0.000673	0.008744	0.006929	0.151537	**0.546308**
#1318	0	1	0	0	0.003006	0.003049	0.003882	0.005230745	0.024152	0.011724	0.015752	0.01966	0.138902	**0.488959**
#1412	0	1	0	0	0.002874	0.006466	0.00431	0.000718391	0.001437	0.002874	0.001437	0.002155	0.16954	**0.567529**

but eventually most of the related tweets were located in the USA (i.e., the measured values in bold, shown in Table 6). The advantage of populations of Twitter users in the USA affects the likelihood of a more impacted result in the spatial aspect in this empirical study. However, as mentioned previously, when analyzing the distribution of Twitter messages for event awareness, it is important to consider such a factor regarding the uneven distribution of the users' locations around the world. Table 5 illustrates the top ten time zone of the volume of experimental messages. In the offline experiment, the spatio-temporal information and patterns of tested events (shown in Table 5 and 6) can be taken for further analysis tasks, but they have not been involved in our final relatedness computation.

6.2.4. Discussion

The experimental results shown in the previous section suggest that the method for measuring event relatedness using the multi-classifier approach provides quite reasonable results, significantly

Table 7
Relatedness of event with baseline event "Virginia earthquake" event at various time points (Event ID: #1173, August 24, 2011).

Event ID	Comparison of event relatedness (baseline: #1173) Measuring technique				
	Cosine	Distance	Dice	Jaccard	Human judgment
#1263	0.89821	0.654377	0.896842	0.812977	0.88
#799	0.866565	0.599884	0.864284	0.761004	0.12
#471	0.89165	0.590953	0.875789	0.779025	0.14
#1292	0.159484	0.006097	0.159108	0.08643	0.18
#3115	0.676199	0.420863	0.67497	0.5094	0.16
#5383	0.783955	0.538932	0.77225	0.628996	0.2
#5647	0.801287	0.537809	0.801151	0.668266	0.18
#1021	0.590271	0.381117	0.5775	0.405976	0.14
#5536	0.052788	−0.02095	0.052787	0.027109	0.12
#2590	0.770872	0.50295	0.770784	0.627054	0.18
#2213	0.817967	0.538117	0.816537	0.689956	0.16
#766	0.887364	0.655452	0.886374	0.795935	0.1
#1680	0.509113	0.336259	0.491059	0.325432	0.16
#2309	**0.920054**	**0.709791**	**0.918377**	**0.849073**	**0.9**
#398	0.914282	0.68896	0.913952	0.841538	0.84
#4204	0.715413	0.461145	0.713048	0.55406	0.76
#1235	0.790477	0.478384	0.783733	0.644375	0.8
#3696	0.600693	0.293916	0.59671	0.425222	0.5
#226	0.275885	0.285516	0.180376	0.099128	0.5
#5994	0.776266	0.452308	0.767325	0.622487	0.8
#3983	0.709327	0.367823	0.69845	0.536629	0.74
#4329	0.71706	0.368877	0.704305	0.543573	0.8
#1339	−0.03341	−0.04098	−0.03338	−0.01642	0.1

Table 8
Information of events.

ID	Icon	Event	Start time	End time
#1263	□	Hurricane Irene	Aug 27 18:14:19 CST 2011	Aug 28 14:46:55 CST 2011
#799	X	London Rioting	Aug 09 00:43:33 CST 2011	Aug 09 11:15:18 CST 2011
#471	◇	Amy Winehouse dead	Jul 23 23:19:19 CST 2011	Jul 24 12:52:58 CST 2011
#1292	◇	MTV Video Music Awards (2011)	Aug 28 16:45:59 CST 2011	Aug 29 10:34:32 CST 2011
#3115	◇	Oscar (2011)	Feb 28 02:12:20 CST 2011	Feb 28 17:49:41 CST 2011
#5383	○	Happy Easter (2011)	Apr 23 01:51:07 CST 2011	Apr 23 02:01:49 CST 2011
#5647	X	Obama Says Osama Bin Laden Is Dead	May 02 09:55:08 CST 2011	May 03 02:08:13 CST 2011
#1021	◎	Super bowl (2011)	Feb 06 21:25:54 CST 2011	Feb 07 16:03:46 CST 2011
#5536	◇	Royal Wedding	Apr 28 14:47:56 CST 2011	Apr 30 11:29:29 CST 2011
#2590	◎	Philippine VS Indonesia (2010)	Dec 16 16:37:35 CST 2010	Dec 17 00:40:56 CST 2010
#2213	◎	Indonesia VS Thailand (2010)	Dec 07 16:46:30 CST 2010	Dec 08 00:26:05 CST 2010
#766	○	Halloween (2010)	Oct 30 23:22:28 CST 2010	Nov 01 15:01:53 CST 2010
#1680	○	Thanksgiving Day (2010)	Nov 24 20:18:50 CST 2010	Nov 26 20:39:46 CST 2010
#2309	△	New Zealand Earthquake	Feb 22 07:50:12 CST 2011	Feb 22 18:36:12 CST 2011
#398	△	Pakistan Earthquake	Jan 19 04:33:06 CST 2011	Jan 19 06:03:28 CST 2011
#4204	△	Philippines Earthquake	Mar 21 18:43:44 CST 2011	Mar 21 19:31:25 CST 2011
#1235	△	Chile Earthquake	Feb 12 04:16:27 CST 2011	Feb 12 06:22:00 CST 2011
#3696	△	Japan Earthquake	Mar 11 13:52:21 CST 201	Mar 13 22:42:09 CST 2011
#226	△	Haiti Earthquake	Jan 13 04:07:00 CST 2011	Jan 13 06:10:36 CST 2011
#5994	△	Spain Earthquake	May 12 01:49:15 CST 2011	May 12 04:50:45 CST 2011
#3983	△	Chile Earthquake	Mar 17 06:47:56 CST 2011	Mar 17 06:52:21 CST 2011
#4329	△	Thailand Earthquake	Mar 24 22:00:36 CST 2011	Mar 25 00:20:56 CST 2011
#1339	◇	Grammy Awards (2011)	Feb 14 01:06:02 CST 2011	Feb 15 14:48:07 CST 2011
#1173	△	Virginia Earthquake	Aug 24 01:52:21 CST 2011	Aug 24 10:57:13 CST 2011
#36	◇	Oscars (2012)	Feb 27 06:06:10 CST 2012	Feb 27 12:53:41 CST 2012
#1175	◎	Super Bowl (2012)	Feb 05 21:45:16 CST 2012	Feb 06 13:39:52 CST 2012
#1318	◇	Whitney Houston dead	Feb 12 07:25:58 CST 2012	Feb 13 16:16:40 CST 2012
#1412	◎	Jeremy Lin	Feb 15 10:00:49 CST 2012	Feb 15 14:22:08 CST 2012

offering a comprehensive way for event evaluation. To allow a more clear demonstration, we also provide alternative representations of example measures of event relatedness based on the experiments mentioned above, including Cosine, Distance, Dice, and Jacard methods, as shown in Table 7. In Table 7, we compare the relatedness of event (Event ID: #1173) with other events by the Cosine, Distance, Dice, and Jacard measuring methods. The results shown in have been verified and validated by human judgments, as described in previous section. As the results shown in Table 7, the performance of offline event measures using supervised approach developed in this work is comparatively not as good as the online evaluation, particularly regarding their efficiency and precision of summarized results. For visualizing the experimental results, we interprete the above empirical results and illustrate some examples of event relatedness in Table 8 and Figs. 9–13. Table 8 illustrates the list of measured events and related information. Figs. 9–13 demonstrate some example measuring results of event relatedness using Cosine similarity measures.

As mentioned previously, the goal of the work is attempted to explore a way to compare the relatedness of events through selected topic categories in their simplest semantic representation. The experimental results indicate that the developed platform is sensible for event evaluation, and can be applied to more general applications such as identifying the relatedness of news articles which may be associated with several topic categories or issues. As a result, our system framework allows users feel free to allocate appropriate amounts of classifiers of selected topics into the platform for evaluation, according to the specific requirements of real world scenarios. Besides, by exploiting the potentials of stored social-media messages, this model can also be used to quantitatively estimate the degree of impact of all investigated events (e.g. financial crisis events) in terms of specified domains.

7. Related work

The related techniques used to identify event relatedness can be categorized into two methods. The first one is to detect event evolution patterns, and the other one is the *story link detection* (*SLD*) technique. Event evolution is defined as the transitional development process of related events within the same topic (Yang, Shi, & Wei, 2009). Some researchers have clearly defined the features of events for mining social streams. (Zhao, Mitra, & Chen, 2007) utilized content-based clustering, temporal intensity-based segmentation, and information flow pattern to define an event for identifying event in social text streams. Becker, Naaman, and Gravano (2009) and Becker et al. (2010) proposed several novel techniques for identifying events and their associated social media documents, by combining content, temporal, and local features of the document. Becker, Chen, Iter, Naaman, and Gravano (2011), Becker, Naaman, and Gravano (2011a, 2011b) and Naaman, Becker, and Gravano (2011) utilized temporal features, social features, topical features, and twitter-centric features to separate event and non-event content in twitter messages stream, aiming to utilize these features for cluster or classify events in social messages streams. Zhai, Velivelli, & Yu (2004) proposed a content-based cross-collection mixture model to discover any latent common themes across all collections and summarize the similarity and differences of these collections along each common theme. Mei and Zhai (2005) studies a particular TTM task-based on content and temporal feature to discover and summarize the evolutionary patterns of themes in a text stream. Spiliopoulou, Ntoutsi, Theodoridis, & Schult (2006) proposed the framework MONIC based on content and temporal feature for modeling and tracking of cluster transitions. Lin, Chi, Zhu, Sundaram, & Tseng (2008) proposed FacetNet based on content feature for analyzing communities and their evolutions through a robust unified process. Leskovec (2011) proposed a technique based on content and temporal feature for finding the relationship among users. Cunha et al. (2011) utilized hashtags for content evolution to analyze the relationship among users. Choudhury, Sundaram, John, Seligmann, & Kelliher (2010) combined user-based, topology-based and time features to extract the information diffusion, and proposed a dynamic Bayesian network based framework to predict the information diffusion at a future time slice in Twitter. Lin proposed a novel probabilistic model called TIDE for the joint inference of diffusion and evolution of topics in social communities Lin, Mei, Jiang, Han, & Qi (2011). They integrated the generation of text, the evolution of topics, and social network structure in a unified model which combine topic model and diffusion model for finding the topic diffusion and topic evolution in DBLP and Twitter. Tang utilized a single-pass clustering algorithm and proposed a topic aspect evolution graph model to combine text information, temporal information, and social information for modeling the evolution relationships among events in social communities (Tang & Yang, 2011). Compared with their work which mainly utilized messages on given topics to detect information diffusion and evolution rather than event formulation and evaluation, our work attempts to utilize various event features and formulation approaches to deal with relatedness computation, allowing for combining online event mining and relatedness evaluation tasks. *Story link detection* (*SLD*) is one of TDT tasks proposed by DARPA, and is mainly used to analyze two stories. In our survey, story link detection techniques can be classified into two categories: one is based on vector-based methods and the other one is based on probabilistic-based methods. Vector-based methods mainly utilized tf-idf to weight and utilized similarity measure to judge the similarity of two stories (Brown, 2002; Chen, Farahat, & Brants, 2004; Ferret, 2002; Shah, Croft, & Jensen, 2006; Wang & Li, 2011; Zhang, Wang, & Chen, 2007, 2008; Štajner & Grobelnik, 2009). Probabilistic-based methods mainly utilized probabilistic model to represent the relationship among words and documents, and utilized many kind of similarity function to measure the association among documents (Nallapati, 2003; Nallapati & Allan, 2002; Nomoto, 2010; Lavrenko, 2002). Story link detection mainly focused on event similarity rather than event evolution (Tang & Yang, 2011; Yang et al., 2009), thus we do not utilize SLD as our approach in this work.

Besides, the data stream clustering has been an important issue in string mining community in recent years (Gaber, Zaslavsky, & Krishnaswamy, 2005). Guha proposed a data stream clustering technique STREAM (Guha, 2003). The *k*-median clustering algorithm was adopted with a simple algorithm based on divide-and-conquer to solve the memory limitation problem. In addition, a stream clustering approach called CluStream that generates an online component which periodically stores detailed summary statistics and an offline component which uses only this summary statistics (Aggarwal, Han, Wang, & Yu, 2003). Zhong combined online spherical *k*-means (OSKM) algorithm with an existing scalable clustering strategy to achieve fast and adaptive clustering of text streams (Zhong, 2005a, 2005b).

In order to deal with online processing data with temporal information, forgetting (half-life) 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.

In our survey most data stream clustering work use *k*-means techniques as their major data 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, especially in dealing with the topic detection task with dynamic topics. Such issues were discussed in Zhong (2005b) 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 density based clustering for extracting event topics from microblogging data collection for relatedness computation.

8. Conclusion

In order to prevent people's lives and properties from being seriously damaged by the unexpected emerging events, it would be helpful to learn the patterns of event evolution from past experiences. In this work, we have utilized Twitter streams to develop a solution combining an online event evaluation system using an unsupervised event clustering approach, and offline measure metrics for comparing relatedness of past events using a supervised SVM-classifier based vector approach. Each of these two models can work independently as a stand-alone system. Once some emerging events were detected by our system, the event clusters and event vectors can be generated by formulating clustered messages by our algorithm. Also, a relatedness measure metrics developed for computing event relatedness can be used for event evaluation. Several essential features of each detected event dataset have been extracted by performing content mining for content analysis, spatial analysis, and temporal analysis. This allows our approach compare the new event vector with other event vectors for evaluation of event relatedness, by means of validating event feature factors involved in the event evolution. The experimental results show that our proposed approach has the potential for online evaluation of related events, and being able to dynamically compare the relatedness among the on-going event with other ones.

On the other hand, in order to establish an offline learned model to quantitatively measure the relatedness of event impacts, we study on analyzing the contents of Twitter datasets related to the events. In this work, we focused on the generation of an extensible model combining eight well-trained SVM classifiers by means of training with the Twitter messages of selected topic domains. The topic classifiers were well developed based upon the best results performed by the training and testing process. Subsequently, we started to perform relatedness evaluation on the platform of the multiple classifiers, and the event datasets were being formulated and go through the categorical decision process of the classifiers to produce formulated event-vectors. The resulting vectors represent the essential features of the respective events, based on the judgments of employed multiple classifiers. Finally, we measured the relatedness of event vectors through the Distance, Cosine, Dice, and Jacard measuring methods and perform a system evaluation by comparing them with human judgments.

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

- In the experiments, our online unsupervised method for evaluation of related events did provide a quick and appropriate result for identification of event relatedness in near real-time. On the other hand, the performance of offline event comparison using supervised measuring approach developed in this work is comparatively not as good as the online evaluation method, in terms of efficiency and precision of summarized results. However, the later one established a framework of the supervised classifier based vector approach for computing the event impact for possible aspects is of great potentials for numerous applicable fields, such as impact evaluation on a financial crisis event. This is due to such a model offers a sensible way to quantitatively analyze the impact factors of all investigated events in selected specific topics, by exploiting the potentials of collected social-media messages.
- In dealing with user generated content in microblogs, a challenging language issue found in messages is in the informal English domain (with no controlled vocabulary), such as abbreviations, named entities, slang and context specific terms in the content; lacking in sufficient context to grammar and spelling. This increases the difficulties in semantic analysis of microblogs. In particular, the length of each message leads to a problem with the lack of semantic integrality in tweets. This makes it fairly difficult to design a reliable weighting and clustering algorithms. In this work, we overcome such challenges, by utilizing our developed dynamic weighting method and clustering algorithm. The preliminary results show that our algorithmic model has the potential for event mining and evaluation.
- In our offline event evaluation model, we have established a comprehensive solution using a model of extensible measure metrics covering several factors to estimate the impacts of studied events on specific domains. By visualizing the experimental results of learned models, we expect to extend the relatedness evaluation model to more separate topic-categories, resulting in an interpretable set of event impacts in more specific aspects. This provides a quantitative investigation of the effect of important factors for historical events, pursuing a deep understanding of known events and their possible inter-relationship. In this work, the results of these set of experiments in relatedness evaluation of events have not been not as good as the results of our online event evaluation method. However, the offline measuring model did provide possibilities for a deeper analysis of event impacts on various domains.

In our future work, we will mainly focus on conducting a detailed study on evaluating other candidate on-line clustering and evaluation methods to carry out microblogging-stream mining, and compare their performance with our developed density-based methods. On the other hand, the second task is to study on developing other unsupervised and supervised learning methods for evaluation and prediction of event development. In the third task, the utilization of advanced name-entity recognition (NER) techniques would be helpful, and can be applied in our system for enhancing event evaluation tasks.

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, Berlin, Germany* (Vol. 29).

Becker, H., Naaman, M., & Gravano, L. (2009). Event identification in social media. In *Proceedings of the ACM SIGMOD workshop on the web and databases (WebDB '09)*.

Becker, H., Naaman, M., & Gravano, L. (2010). Learning similarity metrics for event identification in social media. In *Proceedings of the 3rd ACM International Conference on Web search and data mining, New York, USA*.

Becker, H., Chen, F., Iter, D., Naaman, M., & Gravano, L. (2011). Selecting quality twitter content for events. In *Proceedings of the 25th ACM AAAI international conference on association for the advancement of artificial intelligence, San Francisco, USA*.

Becker, H., Naaman, M., & Gravano, L. (2011a). Beyond trending topics: real-world event identification on Twitter. In *Proceedings of the 25th ACM AAAI international conference on association for the advancement of artificial intelligence, San Francisco, USA*.

Becker, H., Naaman, M., & Gravano, L. (2011b). Automatic identification and presentation of Twitter content for planned events. In *Proceedings of the 25th ACM AAAI international conference on association for the advancement of artificial intelligence, San Francisco, USA*.

Brown, R. D. (2002). Dynamic stopwording for story link detection. In *Proceedings of the 2nd international conference on human language technology research, San Diego, California*.

Chang, C. C., & Lin, C.J . (2001). LIBSVM: A library for support vector machines. In *ACM transactions on intelligent systems and technology (ACM TIST)*.

Chen, F., Farahat, A., & Brants, T. (2004). Multiple similarity measures and source-pair information in story link detection. In *Proceedings of the international conference on human language technology conference of the North American Chapter of the Association for Computational Linguistics, Boston, Massachusetts, USA*.

Choudhury, M. D., Sundaram, H., John, A., Seligmann, D. D., & Kelliher, A. (2010). Birds of a feather: Does user homophily impact information diffusion in social media? In *Proceedings of the computing research repository*.

Cunha, E., Magno, G., Comarela, G., Almeida, V., Gonçalves, M. A., & Benevenuto, F. (2011). Analyzing the dynamic evolution of Hashtags on Twitter: A language-based approach. In *Proceedings of the workshop on languages in social media, Portland, Oregon*.

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 databases.* New York, USA.

Ferret, O. (2002). Using collocations for topic segmentation and link detection. In *Proceedings of the 19th international conference on computational linguistics, Taipei, Taiwan* (Vol. 1).

Gaber, M. M., Zaslavsky, A., & Krishnaswamy, S. (2005). Mining data streams: A review. In *Proceedings of the SIGMOD Records* (Vol. 34, pp. 18–26).

Guha, S. et al. (2003). Clustering data streams: Theory and practice. In *Proceedings of the IEEE transactions on knowledge and data engineering* (Vol. 15, pp. 515–528).

Ishikawa, Y., Chen, Y., & Kitagawa, H., (2001) An on-line document clustering method based on forgetting factors. In *Proceedings of the 5th European conference on research and advanced technology for digital libraries* (pp. 325–339).

Lavrenko, V. et al. (2002). Relevance models for topic detection and tracking. In *Proceedings of the 2nd international conference on human language technology research, San Diego, California, USA*.

Lee, C. H. (2012). *Mining spatio-temporal information on microblogging streams using a density-based online clustering method*. Expert Systems with Applications.

Lee, C. H., Wu, C. H., & Chien, T. F. (2011). BursT: A dynamic term weighting scheme for mining microblogging messages. In *Proceedings of the 8th international symposium in neural networks. Lecture notes in computer science* (Vol. 6677, pp. 548–557). Berlin/Heidelberg: Springer.

Lee, C. H., Yang, H. C., Chien, T. F., & Wen, W.S. (2011). A novel approach for event detection by mining spatio-temporal information on microblogs. In *Proceedings of the IEEE international conference on advances in social network analysis and mining, Kaohsiung, Taiwan, July 25–27*.

Leskovec, J. (2011). Social media analytics: Tracking, modeling and predicting the flow of information through networks. In *Proceedings of the 20th ACM WWW international conference on World Wide Web, Hyderabad, India*.

Lin, Y. R., Chi, Y., Zhu, S., Sundaram, H., & Tseng, B. L. (2008). Facetnet: A framework for analyzing communities and their evolutions in dynamic networks. In *Proceedings of the 17th ACM WWW international conference on World Wide Web, Beijing, China*.

Lin, C. X., Mei, Q., Jiang, Y., Han, J., & Qi, S. (2011). Inferring the diffusion and evolution of topics in social communities. In *Proceedings of the 5th ACM SNAKDD international workshop on social network mining and analysis, San Diego, CA, USA*.

Mei, Q., & Zhai, C. (2005). Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In *Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery in data mining, Chicago, Illinois, 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*.

Nallapati, R. (2003). Semantic language models for topic detection and tracking. In *Proceedings of the international conference on the North American Chapter of the Association for Computational Linguistics on Human Language Technology: HLT-NAACL 2003 student research workshop, Edmonton, Canada* (Vol. 3).

Nallapati, R., & Allan, J. (2002). Capturing term dependencies using a language model based on sentence trees. In *Proceedings of the 8th international conference on information and knowledge management, McLean, Virginia, USA*.

Nomoto, T. (2010). Two-Tier similarity model for story link detection. In *Proceedings of the 19th ACM international conference on information and knowledge management, Toronto, ON, Canada*.

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).

Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc.

Shah, C., Croft, W. B., & Jensen, D. (2006). Representing documents with named entities for story link detection (SLD). In *Proceedings of the 15th ACM international conference on information and knowledge management, Arlington, Virginia, USA*.

Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y., & Schult, R. (2006). MONIC: Modeling and monitoring cluster transitions. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, Philadelphia, PA, USA*.

Štajner, T., & Grobelnik, M. (2009). Story link detection with entity resolution. In *Proceedings of the 8th ACM WWW international conference on World Wide Web semantic search workshop, Madrid, Spain*.

Tang, X., & Yang, C. C. (2011). Following the social media: Aspect evolution of online discussion. In *Proceedings of the 4th international conference on social computing, behavioral-cultural modeling and prediction, College Park, USA*.

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* (pp. 208–215).

Vapnik, V. (1999). An overview of statistical learning theory. *IEEE Transaction on Neural Networks, 10*(5), 988–999.

Wang, L., & Li, F. (2011). Story link detection based on event words. In *Proceedings of the 12th international conference on computational linguistics and intelligent text processing, Tokyo, Japan* (Vol. Part II).

Yang, C. C., Shi, X., & Wei, C. P. (2009). Discovering event evolution graphs from news corpora. *Proceedings of the IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 39*, 850–863.

Zhai, C., Velivelli, A., & Yu, B. (2004). A cross-collection mixture model for comparative text mining. In *Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining, Seattle, WA, USA*.

Zhang, X., Wang, T., & Chen, H. (2007). Story link detection based on dynamic information extending. In *Proceedings of the 45th annual meeting of the association for computational linguistics, Prague, Czech Republic*.

Zhang, X., Wang, T., & Chen, H. (2008). Story link detection based on event model with uneven SVM. In *Proceedings of the 4th Asia information retrieval conference on information retrieval technology, Harbin, China*.

Zhao, Q., Mitra, P., & Chen, B. (2007). Temporal and information flow based event detection from social text streams. In Proceedings of the 22nd international conference on artificial intelligence, Vancouver, Canada (Vol. 2).

Zhong, S. (2005a). Efficient streaming text clustering. *Proceedings of the Neural Networks, 18*, 790–798.

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