Soc. Netw. Anal. Min. (2016) 6:107
DOI 10.1007/s13278-016-0414-1

REVIEW ARTICLE

A survey of event detection techniques in online social networks

Anuradha Goswami[1] · Ajey Kumar[1]

Received: 10 May 2016 / Revised: 20 October 2016 / Accepted: 25 October 2016 / Published online: 17 November 2016
© Springer-Verlag Wien 2016

Abstract The online social networks (OSNs) have become an important platform for detecting real-world event in recent years. These real-world events are detected by analyzing huge social-stream data available on different OSN platforms. Event detection has become significant because it contains substantial information which describes different scenarios during events or crisis. This information further helps to enable contextual decision making, regarding the event location, content and the temporal specifications. Several studies exist, which offers plethora of frameworks and tools for detecting and analyzing events used for applications like crisis management, monitoring and predicting events in different OSN platforms. In this paper, a survey is done for event detection techniques in OSN based on social text streams—newswire, web forums, emails, blogs and microblogs, for natural disasters, trending or emerging topics and public opinion-based events. The work done and the open problems are explicitly mentioned for each social stream. Further, this paper elucidates the list of event detection tools available for the researchers.

Keywords OSN · Event detection · Social text streams · Thematic · Temporal · Spatial · Network structure · Natural disaster · Emerging topics · Public opinion

✉ Ajey Kumar
ajeykumar@scit.edu

Anuradha Goswami
anuradha@scit.edu

[1] Symbiosis Centre of Information Technology (SCIT), Symbiosis International University (SIU), Rajiv Gandhi Infotech Park, Hinjewadi, Pune, Maharashtra State 411057, India

1 Introduction

Since the inception of Internet in the early 1990s, people are interacting and communicating information in various forms over the web. The era of Web 2.0 further added to this volume of information in the form of World Wide Web content. Online social network (OSN) is a "group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, which allows the creation and exchange of user-generated content" (Kaplan and Haenlein 2010). According to VNI, the global Internet traffic is estimated to grow to 168 exabytes per month by 2019, up from 59.9 exabytes per month in 2014, a 23% in CAGR as shown in Fig. 1 (Cisco 2014). Most of this traffic is contributed by the OSN platforms such as Youtube (Youtube 2016), Facebook (Facebook 2016) and Twitter (Twitter 2016) in the form of either text, videos, images or photographs (Deitrick and Hu 2013). Table 1 shows the data explosion statistics for different OSNs in every minute of the day.

Researchers can get substantial information by churning these explosive data during different kinds of events like earthquake disasters, epidemic disease outbreaks, monitoring traffic control, news events like forest fires and electoral polls (Nurwidyantoro and Winarko 2013). Substantial information could be regarding the people who are influenced by the event, location and time of the event, extent of damage created and effects on surrounding environments. For example, during Queensland floods, data from OSN described the whole situation revealing the yacht sinking on Brisbane River, reopening of the port, incident of bull shark on a flooded street, etc. (Zhou and Chen 2014). The most frequent and recent events which got reflected in OSN and generates massive data or information

Fig. 1 Cisco VNI global Internet traffic prediction

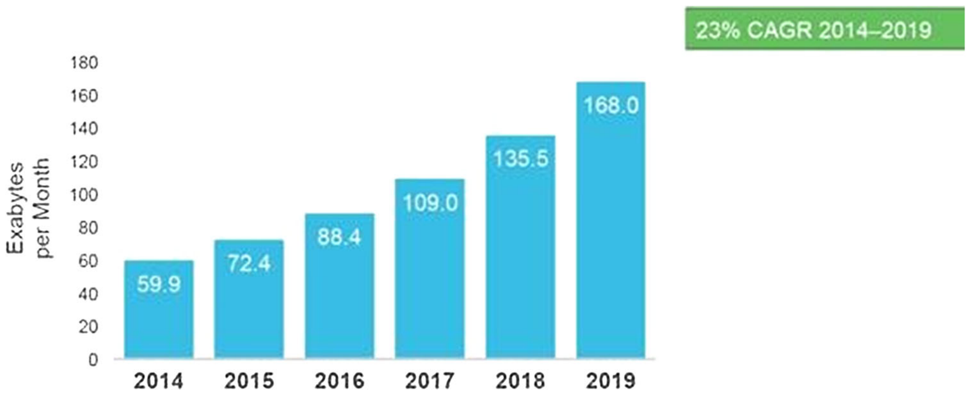


Table 1 Data generation rate for OSN

OSN	Data volume generated/minute of the day
Facebook	2,460,000 pieces of content shared
Twitter	Tweets 277,000 times amounting to 400 million tweets per day
Youtube	Users upload 72 h of new videos
Pinterest	Users pin 3472 images
Instagram	Users posts 216,000 new photographs

Table 2 Event categories

Event category	Examples
Natural or manmade disaster event (NMDE)	Earthquake, tornadoes, floods, epidemics, fire, etc
Emerging event (EE)	World cup, product launch
Public opinion event (POE)	Electoral polls

can be categorized broadly into three groups as shown in Table 2 (Madani et al. 2014).

Event detection has drawn considerable attention of researchers to automatically understand, extract and summarize the happenings of an event in different fields like biosurveillance, safety, health and economics (Tork 2011). In the field of disease outbreaks, DARPA (Neill and Wong 2009) has estimated that a 2-day gain in the detection time can reduce the extent of fatalities by factor of six. Also, studies in (Neill and Wong 2009; Neill and Gorr 2007) have showed that a hot spot violent crime can be detected 1–3 weeks prior to the event by detecting the clusters of leading indication crimes like disorderly conduct and assault. In another study in the field of safety, detection of anomalous clusters of pipe breakage is done supporting the monitoring system of city's water distribution (Neill and Wong 2009). Environmental monitoring is done through remote sensors which are used for continuous observation of a certain place or domain, generating large volumes of data to analyze the same (Dereszynski and Dietterich 2007). So it is evident that event detection in OSN contributes much to understanding or predicting the events. Though identification of emerging events and getting important insights from the same is very important, there are challenges in which these event detection techniques confront in general (Kerman et al. 2009; Neill and Wong 2009; Tork 2011; Madani et al. 2014).

- *Domain dependence:* An event detection procedure or technique which is suitable for one domain might not be the same for the other domains, and it is extremely situational dependent (Neill and Wong 2009). For example, the selection of parameter, variables and output metrics for predicting the electoral poll results will not be the same as the prediction of any natural disaster event, i.e., earthquake.
- *Time Constraint:* An extreme timeline constraint is a timeline in which the event detection method should be able to identify event correctly. Depending upon the domain criticality, the timeline can range from seconds to several minutes. For example, detection of any terrorist attack or measuring indicators of imminent catastrophic disasters are critical applications where constraint on a specific timeline is an important consideration.

- *Detection accuracy:* A high degree of precision is required to be maintained for achieving event detection accuracy in critical domains. Generation of alarm for a true event in mission critical domains like healthcare and banking sectors is very crucial. A false alarm generation due to inaccurate detection of event could involve huge monetary loss and should be considered as a serious challenge. High precision can be maintained by event detection methods through providing a high true-positive rate (accurate detection) and at the same time ensuring a low false-positive rate (inaccurate detection) for events (Margineantu et al. 2010).
- *Diversified data sources:* OSN has effectively contributed to huge explosion of diversified data consisting of unstructured data, textual documents, images, audio, video, relational data, multivariate records and spatiotemporal data (Kerman et al. 2009). Thus, event detection problem is encountered with determining what data are relevant for the event detection under study and the approach which must be opted to evaluate the data from selected sources.
- *Voluminous data:* Huge volume of data requires high-powered computing algorithm and immense storage space to store, access, filter and process all data within stipulated time frame. For example, millions of tweets get generated each day in the Twitter platform. So, to process these huge data against some particular event, event detection algorithms should incorporate dynamism and suitable running environment so that it runs uninterruptedly even after sudden voluminous increase in the OSN data during some bursty events (Li et al. 2014).
- *Authenticity and missing data:* The event detection techniques should consider the inaccuracies and incompleteness of the raw sensor data (Balazinska 2007). For example, the position or location information in terms of longitude and latitude is very likely to be inaccurate or missing. The environmental activity information may have limited confidence level. So, event detection algorithms should consider this underlying incompleteness, inaccuracies and confidence levels while detecting the events in OSNs.
- *Handling anomalous behavior:* Historical data from OSN consist of the mixture of both normal and anomalous event data mixed together (Ihler et al. 2006). The event detection method should first learn the predictable behavior pattern of the event as well as should be capable of detecting and extracting deviated patterns from the raw data. This can be applied for predicting the number of customers entering a bank or predicting the number of freeway traffic accidents per day.

The primary motivation of event detection analysis study got initiated by a project called Topic Detection and Tracking which was a joint venture of DARPA, CMU and Dragon systems (Papka and Allan 1998). According to one of the issues of this project, event is distinguished from the term topic by the property of time. For example, “Mumbai Terrorist Attack on 26 November, 2008” (CNN Library 2015) is considered to be an event, but “terrorist attacks” is a more general topic. This difference can be extended by incorporating the spatial or location aspect in the definition (Papka and Allan 1998).

According to Zhou and Chen (2014), any input message by a user posted to an OSN can be considered as his observation on a real-world happening at a certain location and time. This observation is an *event* element. Given a set of event elements say E, an event is defined as a subset E_i of E, provided all the event elements in the set E_i refer to the same real-world occurrence. Any event can be described by the tuple given as follows:

$$E_i = \{\mathcal{D}_i, \langle la_i, lo_i \rangle, t_i, D_i\}$$

where ‘i’ same real-world occurrence

In the above tuple, $\mathcal{D}_i$ represents data or textual content, $\langle la_i, lo_i \rangle$ represents location in terms of latitude and longitude of the place of the event, t_i denotes timestamp attached to the message indicating the approximate time of occurrence of the event, and D_i represents the social links connecting the users and their followers forming the network structure associated with the flow of the event data (Jadhav et al. 2010; Zhao and Mitra 2007; Zhou and Chen 2014) as shown in Table 3.

The main challenge in event detection for the researchers is to select appropriate tools and techniques while detecting events in OSN which would be in sync with the type of event and event detection methods discussed above. This motivates us to focus on the existing tools and techniques available in event detection for OSN. In this paper, three types of events will be considered: natural disasters, trending or emerging topics and public opinion-based events. The survey of event detection techniques used for the above-mentioned three types of events will be done for social text stream data (described in Sect. 2), and the open challenges will be discussed. Tools available for doing event detections will also be explained with the aim of its applicability for the researchers.

The remaining of this paper is organized as follows. Section 2 reviews the existing event detection techniques for different types of social text stream data, and open challenges for the researchers are discussed. In Sect. 3, comparison of related tools available for doing event detection is done. Finally, conclusion is given in Sect. 4.

Table 3 Event detection techniques based on event dimensions

Event dimensions	Definition	Event detection techniques
Thematic (TH)	Techniques used to detect events exploiting the semantic and contextual features of the data or textual content ($\mathcal{D}_i$)	Term and named entity vector (Kumaran and Allan 2004) Clustering algorithms (Aggarwal et al. 2012) Incremental greedy clustering algorithms (Allan et al. 1998) Gaussian mixture models (He et al. 2007a, b) Binomial distribution (Fung et al. 2005) Discrete Fourier transformation (He et al. 2007a, b) Naive Bayes classifier (Sankaranarayanan et al. 2009) Cosine similarity (Petrovic et al. 2010) Support vector machine (SVM) (Becker et al. 2011) Latent Dirichlet allocation (Blei et al. 2003) Gradient boosted decision trees (Friedman 2001) Conditional random field (CRF) (Benson et al. 2011) ETree using n-gram-based content analysis techniques (Gu et al. 2011) Gradient boosted decision trees (Popescu and Pennacchiotti 2010; Popescu et al. 2011)
Temporal (T)	Techniques used for detecting events considering the timestamp (t_i) of the message ($\mathcal{D}_i$) through the variation of number of messages over a specific time period	Wavelet transformation and normalized wavelet entropy (Weng and Lee 2011) TSCAN using eigenvectors of a temporal block association matrix (Chen and Chen 2008) Temporal and dynamic query expansion techniques (Massoudi et al. 2011; Metzler et al. 2012) Spectral analysis and weighted graph (Cordeiro 2012, Long et al. 2011, Weng and Lee 2011) Hidden Markov models (Khreich et al. 2012) Multivariate event detection methods (Neill and Wong 2009)
Spatial (S)	Techniques which use the location information in terms of latitude and longitude $\langle la_i, lo_i \rangle$ of a message ($\mathcal{D}_i$) to detect events	Locality-sensitive hashing methods (Gionis et al. 1999) Kalman and particle filtering (Fox et al. 2003) Univariate spatial scan statistic approaches (Neill and Wong 2009) Multivariate Bayesian spatial scan statistic methods (Neill and Wong 2009) Parametric scan statistic methods (Neill and Wong 2009) Nonparametric scan statistic methods (Neill and Wong 2009) Spatial–temporal random indexing (Menon et al. 2010)
Network structure (NS)	Techniques which study the social information (D_i) through the embedded network structure or the information flow pattern between users and the user connections of OSN for event detection	Dynamic time warping concept (Keogh 2002) Graph cut algorithm (Shi and Malik 2000)

2 Social text stream-based event detection techniques

Social text stream data are a collection of text communication data, where each text stream is associated with attributes such as *author/sender and reviewer/recipients* (Zhao et al. 2007). The authors/senders are the creator of the message or data, and reviewer/recipients are the viewer or receiver of the same message or data. This message is mostly about real-world events information which is semantically significant and generated from various types of sources (Kleinberg 2006). The primary sources or applications which generate these stream data are newswires, web forums, emails and blogs or microblogs from social networking sites as shown in Fig. 2 (Zhao et al. 2007; Chen et al. 2012).

Event detection is a mechanism through which events are identified automatically from OSN data which consist of information regarding what happened, when, where, to whom and why (Zhou et al. 2014). These detection tasks require high level of data throughput accompanied by a low degree of response latency (McCreadie et al. 2013). There

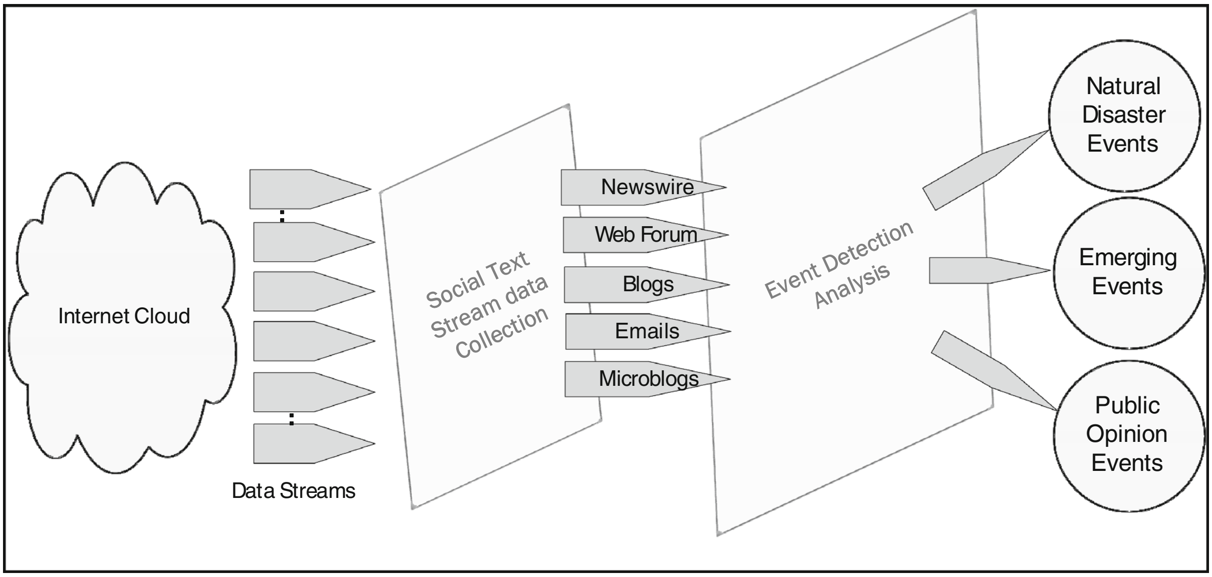


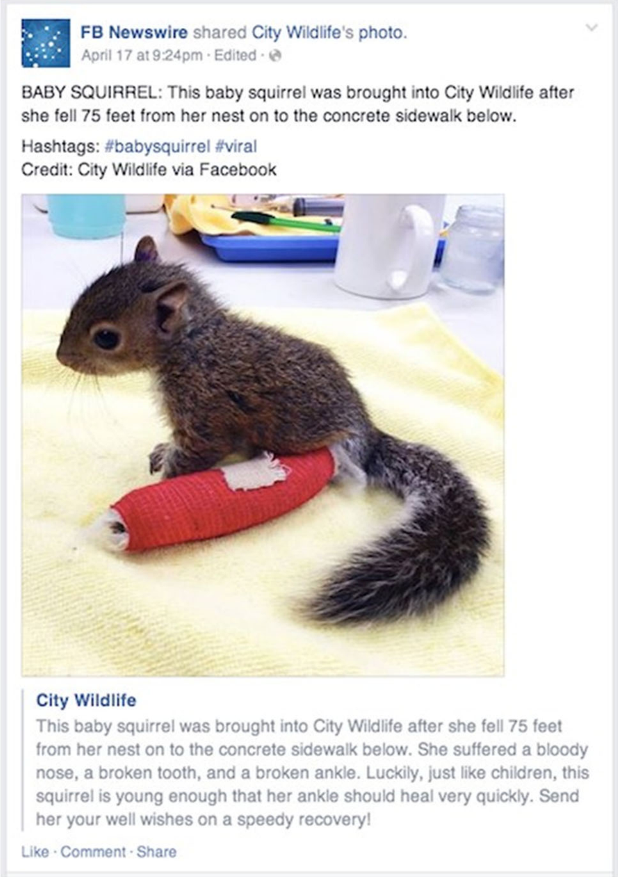
Fig. 2 Social text stream data

are existing surveys which are done in general on OSN (Nurwidyantoro and Winarko 2013) and Twitter in specific (Atefeh and Khreich 2015; Zhou and Chen 2014). The latest survey done on Twitter streams by Atefeh and Khreich (2015) has classified the existing event detection techniques on Twitter streams with respect to the event type (specified or unspecified), type of detection task (retrospective or new event detection), methods of detection (supervised and unsupervised) and also the applications for which the techniques are applied. This section reviews the existing event detection studies across different types of OSN streams, viz. newswire, web forums, blogs, emails and microblogs. The event detection techniques proposed are categorized exploiting the content flow between users (thematic), the temporal information (temporal), location (spatial) and social structure (network structure) details from the OSN data (Metzler et al. 2005; Qi and Candan 2006; Song et al. 2006). This section reviews the existing studies on event detection techniques for different types of social text stream data.

2.1 Newswires

2.1.1 Description

Newswires are services which transfers latest news stories via satellite or Internet (Newswires 2015). Event detection was done as an extension to text retrieval and clustering techniques to automatically detect new events from newswires. For example, as shown in Fig. 3, Facebook Newswire launched by Facebook with News Corp-owned Storyful provides repository of verified, real-time content which covers breaking news stories. These contents are mainly shared publicly by the Facebook users in the form of text or video or photographs on account of some breaking news.



FB Newswire shared City Wildlife's photo.
April 17 at 9:24pm · Edited ·

BABY SQUIRREL: This baby squirrel was brought into City Wildlife after she fell 75 feet from her nest on to the concrete sidewalk below.

Hashtags: #babysquirrel #viral
Credit: City Wildlife via Facebook

City Wildlife
This baby squirrel was brought into City Wildlife after she fell 75 feet from her nest on to the concrete sidewalk below. She suffered a bloody nose, a broken tooth, and a broken ankle. Luckily, just like children, this squirrel is young enough that her ankle should heal very quickly. Send her your well wishes on a speedy recovery!

Like · Comment · Share

Fig. 3 Facebook newswire (*source:* www.poynter.org)

2.1.2 Event detection techniques

The studies of techniques for event detection are categorized according to the differentiation of methods in Table 3. Few studies exist where there is a mixture of techniques between categories from different event detection methods.

Traditionally, Yang et al. (1998) used newswires to automatically detect retrospective and online events exploiting the ordered temporal patterns of streams of news stories from CNN news and Reuter's articles. The study applied an agglomerative method called group average clustering (GAC) (Yang et al. 1998) to obtain clusters of news stories having temporal proximity in the given event. A single-pass incremental clustering method (Yang et al. 1998) is used to handle the dynamic nature of streaming data and the temporal properties of the events. The study showed an improvement in the event cluster quality detection by incorporating the content and temporal features of the data streams. The study in (Yang et al. 1998) was extended by Yang et al. (1999) which dealt with the same approach to event detection and additionally taken up event tracking in the same mentioned newswire corpus. Event tracking was done as supervised learning and aimed to automatically assign event levels to the incoming news stories in the data streams depending on the past learning of events on a small set of past stories identified. They showed a quick learning is desired to distinguish between different topically related events and is achieved by maintaining a small number of positive training examples per event to track an event. Machine learning algorithms such as k-nearest neighbor classification (Yang et al. 1999) and decision-tree induction method (Yang et al. 1999) were used for event tracking. Another similar study in (Allan et al. 1998) used newswires and transcribed broadcast news in detecting only new online events. The single-pass clustering algorithm was modified in this study to deal with the online news stories sequentially as a data stream. A thresholding model is used to exploit the time proximity of two news stories on the stream where chances of two stories discussing the same event are more if the proximity is low. A latest study by Schubotz and Krestel (2015) was done on the online temporal summarization. A temporal summarizer was developed against a given query by the user which is based on probabilistic language models and entity recognition. BM25 scoring (Schubotz and Krestel 2015) is used to extract all relevant streams of news from the newswires. A general query language model with Kullback–Leibler divergence (Schubotz and Krestel 2015) is used to detect sentences matching the query from the relevant streams. Lastly, the novel sentences which constitute the new events are extracted for summarization of the query from the relevant sentences. All the above studies exploited the temporal as well as the thematic features of newswires.

A novel approach was suggested by Menon et al. (2010) for automatic event detection in newswire data, using spatial–temporal random indexing. The study involves a 4-dimensional vector space where each word in the corpus is specifically represented by a large sparse vector, unique in time and location. This spatially and temporally annotated space is analyzed to find significant semantic shifts over time and space. This shift is finally used to predict the occurrence of events. Automatic new event detection was also done by Lam et al. (2001) by locating events related topically in a continuous stream of chronological multilingual newswire stories. This study used concept terms, named entities and story terms to represent a story or event which improved the detection task. Event is detected through concept term extraction from the concept database against a sentential match. Part-of-speech tagger is used to extract named entities from the news story. The Chinese stories are translated and converted into a set of English words on which agglomerative clustering was applied to obtain clusters of related stories. This study successfully cross-validated topically related events in two different multilingual streams and also used natural language processing for event detection.

Many event detection studies involved the use of ACE (automatic content extraction) 2005 corpus of newswires. This corpus consists of a set of 33 generic event types and subtypes appearing frequently in news (Ahn 2006; Grishman et al. 2005). Few terminologies which got introduced through ACE needs to be elaborated to proceed with further studies in this arena (Hong et al. 2011). The terminologies are:

- *Entity*: An object or a set of objects in any one of the semantic categories of interest, referred in a document by one or more co-referential event mentions.
- *Entity mention*: A reference to an entity, typically a noun phrase.
- *Entity trigger*: The main word, typically a verb or a noun, that most clearly expresses the occurrence of an event.
- *Event arguments*: The entity mentions, viz. participants who are involved in the event.
- *Argument roles*: The relation of arguments to the event where they participate.
- *Event mention*: A phrase or sentence in which the event is mentioned, including trigger and arguments.

In the study by Ahn (2006), ACE 2005 corpus was used where the entities, the temporal factor and a limited set of values like contact information and job titles of the entities were used for event detection. The study extracted events through a series of classification of subtasks, each of which is handled by a machine-learned classifier. A similar study by Hardy et al. (2006), worked on the thematic features of the set of 37,444 documents from the Center for Nonproliferation Studies corpus, used event-based, data-driven thematic processing and natural language processing accompanied by a visualization interface and contextual information to give appropriate answers to queries. The logistic regression model, Naive Bayes classifier, random error pruning tree are some of the machine learning algorithms used by this study. Another study by Maslennikov and Chua (2007) exploited the thematic features at a sentence level of newswires through a multiresolution framework for sentence-level phrases and clauses using dependency and discourse relations. Discourse analysis deals with the clauses in a sentence connected by clausal relations (Halliday and Hasan 1976). The proposed framework uses the clausal relations to filter the noisy dependency paths and to enhance the reliability of the extraction process of similar paths. A study by Ji and Grishman (2008) extended the scope of discourse relation from one single sentence of one document to a cluster of topically related documents. This study employed ACE event extraction corpus to combine the global features of the documents from the individual local features of each document. Maximum entropy-based classifiers were used to extract events from the corpus. This study can generate an event-driven summary through document clustering and text summarization to enhance support to the users. Discourse analysis was also used in another study by Patwardhan and Riloff (2009) to do a deep structural analysis for associating factual data with event roles. This study considered both local context around a phrase and the more general sentential context for event extraction. A probabilistic framework was used to make a joint decision based on both local evidence from each phrase and general “peripheral vision” from the sentential event recognizer. The framework was supported by machine learning classifier called Naive Bayes classifier (Patwardhan and Riloff 2009) for both at the phrase and the sentential level. Another similar study by Liao and Grishman (2010) involved document-level information for event detection from ACE event extraction corpus. They incorporated more than one event information to make predictions regarding the occurrence of an event. Maximum entropy-based classifier is used to attain a document-level within-event and cross-event consistency. Most of the above studies dealt with the thematic methods to detect events from newswire corpus.

Joint frameworks using entities along with event mentions also boosted the performance of event detection in newswire corpus. A study by Hong et al. (2011) exploiting transductive inference is cross-entity inference which used the relationship between entities exploiting the network/social structure of the entities for event extraction from ACE corpus. They hypothesize that entities of similar events, playing the same role, are normally consistent. A clustering technique called CLUTO toolkit (Hong et al. 2011) is used to generate clusters of sentences or event mentions having entities of the same background. A support vector machine (SVM) classifier (Joachims 1998) is used to distinguish arguments with nonarguments w.r.t roles and entity types to determine the reportable event mention.

A study by Kastner and Monz (2009) focused on extracting only important facts or sentences from newswire article. The identification was done exploiting the syntactic, semantic and statistical features of the sentences in a document. Some of the linguistic features the study used ranges from n-gram frequency through the sentence position to the semantic analysis including spawned phrases, verb classes and types of adverbs. Another similar study (Dasigi and Hovy 2014) introduced a novel technique to model events based on recursive neural networks (Goller and Kuchler 1996) which was trained to differentiate between normal and anomalous events. Anomaly was defined as *unexpected or unusual combination of semantic role fillers, and anomalous events are defined as semantically coherent, but unusual only based on real-world knowledge*. This study pointed out that set of irrelevant or meaningless events and good normal events are two extremes of the semantic spectrum range where anomalous events lies somewhere in between.

Most of the traditional approaches relied on sequential pipelines having multiple stages for event extraction where prediction of event triggers and arguments is done independently (Hong et al. 2011). A study by Li et al. (2013) proposed a joint framework by incorporating the global features which captures the dependencies of multiple triggers and arguments for the ACE corpus. They used a structured perceptron to train this joint framework. An interdomain event extraction method was studied by Miwa et al. (2014) where a system designed for biomedical domain was adapted to work for newswire domain. The study addressed the major differences in the tasks of event detection in the two domains and showed it is needless to develop a separate system for event extraction for different domains.

Table 4 summarizes the techniques mentioned above in context of events detection from different newswires sources. The techniques are categorized according to the type of event dimensions (S—spatial, T—temporal, TH—thematic, NS—network structure) used by the researchers.

2.1.3 Open problems

- *Requirement of techniques to increase user involvement through the temporal clusters at different granularity levels of the newswires documents:* During any disaster event, the users might intend to know details of the event by exploiting temporal patterns based on proper names and proximity phrases leading to the formation and

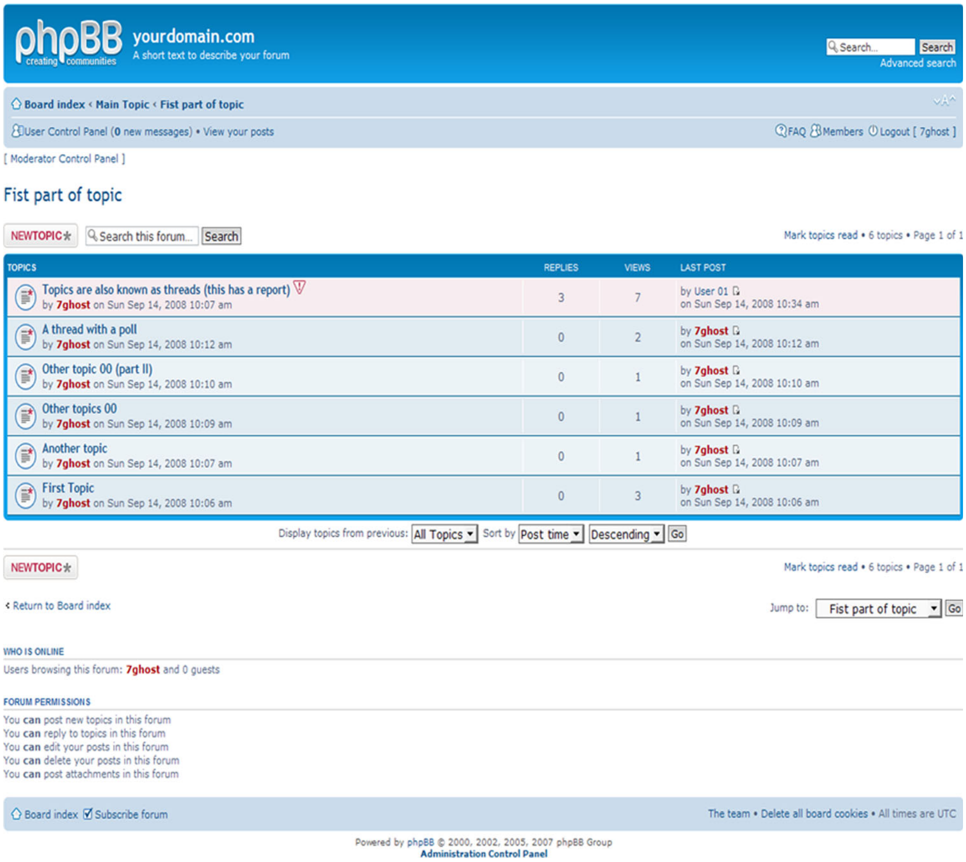
Table 4 List of event detection techniques for newswires

Type of streams	Dimension				Techniques used
	S	T	TH	NS	
Newswires		✔	✔		Group average clustering
		✔	✔		Single-pass clustering
		✔	✔		K-nearest neighbor
		✔	✔		Decision tree
		✔	✔		Thresholding model
		✔	✔		BM25 scoring
			✔		Kullback–Leibler divergence
	✔	✔	✔		Spatiotemporal random indexing
			✔		Part-of-speech tagger
			✔		Logistic regression model
			✔		Naive Bayes classifier
			✔		Random error pruning
			✔		Maximum entropy-based classifier
			✔		Support vector machine (SVM)
			✔		n-gram frequency
			✔		Recursive neural networks
			✔		Structured perceptron

development of temporal clusters of missing people, depicting the damage, etc., across the entire timeline of the event. These temporal clusters can be demanded at the corpus level, document level or subdocument level by the user's query depending upon their requirement during a disaster event. Better techniques are required to generate these temporal clusters at different granular levels by the researchers to create a better knowledge base about the event through newswires.

- *Methods to assist the users in searching relevant cluster at a minimum time:* Time plays a crucial role during events for the users in getting proper news from the newswires. The searching time required to find the most relevant temporal cluster or clusters from the newswires is crucial for users. For example, during disaster management, several decision-making activities like most affected locations to decide on missing people list, indicating the extent of damage to provide relief measures, etc., are tightly coupled with the information available regarding the event. Lesser the time taken to get event information from newswire documents, better will be the decision-making process.
- *Techniques to model event evolution over time more accurately other than using time windowing:* Time windowing is a reputed process of detecting events evolution overtime through the semantic shift of words between different time spans or windows in newswires. But as the events generally occur and disappear over a short span, ranging from days to few months, large time window or span cannot detect events accurately for the researchers. So, methods or techniques to detect event evolution accurately other than time windowing are required.
- *Requirement of unsupervised models:* Traditionally, query-based, keyword extraction-based methods were available to facilitate event extraction process. But, there is a requirement to detect unknown events also, as all events do not happen with prior prediction or intimation. So, automation of event extraction from an unknown corpus is required to detect important unknown topical events by analyzing the textual features of the newswire document.
- *Improved techniques required for relation detection and co-reference resolution:* Two *Event Arguments* referring to the same role of related events make it difficult for the researchers to detect the co-referred event. So, novel techniques are required for event detection considering resolving these relation detections between entities of related events and co-refer it as the same event. For example, consider these two statements:
    - Mr Derreck quit as vice president of XYZ Ltd. last week.
    - XYZ Ltd. employees were on strike for last one month.

Here, two different events mention "Personnel End-Position" and "Company-Strike" refer to two arguments of the same Role "Company – XYZ Ltd." In this situation, proper techniques should be there which will detect the relations between different entities of these events and co-refer both the statements to be under one event "XYZ Ltd at the verge of Closure."


phpBB
creating communities
yourdomain.com
A short text to describe your forum
Search...
Search
Advanced search
Board index ‹ Main Topic ‹ Fist part of topic
User Control Panel (0 new messages) • View your posts
FAQ
Members
Logout [7ghost]
[Moderator Control Panel]
NEWTOPIC
Search this forum...
Search
Mark topics read • 6 topics • Page 1 of 1

Fist part of topic

TOPICS	REPLIES	VIEWS	LAST POST
Topics are also known as threads (this has a report) by 7ghost on Sun Sep 14, 2008 10:07 am	3	7	by User 01 on Sun Sep 14, 2008 10:34 am
A thread with a poll by 7ghost on Sun Sep 14, 2008 10:12 am	0	2	by 7ghost on Sun Sep 14, 2008 10:12 am
Other topic 00 (part II) by 7ghost on Sun Sep 14, 2008 10:10 am	0	1	by 7ghost on Sun Sep 14, 2008 10:10 am
Other topics 00 by 7ghost on Sun Sep 14, 2008 10:09 am	0	1	by 7ghost on Sun Sep 14, 2008 10:09 am
Another topic by 7ghost on Sun Sep 14, 2008 10:07 am	0	1	by 7ghost on Sun Sep 14, 2008 10:07 am
First Topic by 7ghost on Sun Sep 14, 2008 10:06 am	0	3	by 7ghost on Sun Sep 14, 2008 10:06 am

Display topics from previous: All Topics Sort by Post time Descending Go

NEWTOPIC
Mark topics read • 6 topics • Page 1 of 1
‹ Return to Board index
Jump to: Fist part of topic Go

WHO IS ONLINE

Users browsing this forum: 7ghost and 0 guests

FORUM PERMISSIONS

You **can** post new topics in this forum
You **can** reply to topics in this forum
You **can** edit your posts in this forum
You **can** delete your posts in this forum
You **can** post attachments in this forum

Board index ☑ Subscribe forum
The team • Delete all board cookies • All times are UTC
Powered by phpBB © 2000, 2002, 2005, 2007 phpBB Group
Administration Control Panel

Fig. 4 Illustration of web forum (*source:* https://en.wikipedia.org/wiki/Internet_forum)

- *Requirement of cross-domain techniques:* General frameworks for event detection techniques which can be used in more than one domain with minor adaptive changes should be researched further. For example, different approaches which got proved in newswire domain for event detection such as use of contextual information at the document level can be experimented in other domain and vice versa. Researchers should also find out the possibilities of discovering information which can be shared between domains for event detection. This will enhance new directions in research for event extraction tasks in newswire domain along with other related domains.

2.2 Web forums

2.2.1 *Description*

Web forums are platforms on Internet and comprised of content generated by users through exchange of information among them (Bamrah et al. 2014). Exchange of information happens through viewing and sending posts over forums. Web forums select a set of jargons related to them which constitutes the subforums. The users can create new topics in these subforums. Few examples of the topics could be entertainment, games, technical discussions, etc. Under every topic, every new discussion initiated by the users is called a *thread*. For a single thread, there can be many reply posts created by as many anonymous or registered users. Figure 4 illustrates the web forum of phpBB where there are five threads to discuss on for the users. A new topic can be initiated through NEWTOPIC button. There is also a consolidated view of the total number of posts and views for each thread and the user last viewed along with the timestamps.

2.2.2 *Event detection techniques*

Web forums have been successfully used as a platform for event detection (Chester et al. 2011). For example, a large campylobacteriosis outbreak happened in 2007 and was associated with mountain bike race in British Columbia (BC), Canada. Later investigation found out that ingestion of contaminated mud was the source of illness (Stuart et al. 2010). A pre-existing web forum for mountain biking in BC facilitated hugely through postings, related discussion threads and photographs for the investigators to understand the conditions of the race course and also extreme end coverage of racer hands and faces (Chester et al. 2011).

There are various studies done on web forums involving different categories of event detection methods. Traditionally, the web or online forums did not gain so much of reputation as an event detection platform because the postings are imprecise, terse and have a casual communication styles (Zhi et al. 2007). The poster's posting characteristics were used for the first time for event detection in (Zhi et al. 2007) by considering the participation frequency of the posters in the web forum instead of the content posted. The study used nonnegative matrix factorization technique to develop user participation model and detect topics, automatic discovery of leaders and subcommunities in the web forum. A similar study by Cheng and Li (2007) considered user participation instead of post content but employed Markov logic network (MLN) (Richardson and Domingos 2006) to find topic clusters by best fitting a set of rules. Both the studies can be highlighted under the network structure category of the event detection methods.

Discussions on a thread over web forum are considered to be thematically similar if large number of users are discussing on that same thread. Similarly, two users are treated similar if they post the same topic and get involved in the same discussions. A study by Zhu et al. (2008) proposed a different set of thematic solutions in addition to the existing solution to address the problems regarding the topic detection tasks in web forums. They introduced a post and thread activity validation step to filter out uninformative contents or noise out of the content data. A term pos-weighting strategy (Zhu et al. 2008) is used to focus on the analysis of the integral part of the data and also the user activities in the forum. A standard one-class SVM classifier is used for the validation step, and incremental TF-IDF model is used for the pos-weighting strategy and UF-ITUF (user frequency—inverse thread user frequency) to model the content similarity among the users in the forum. The combined score of content similarity and the user activity similarity decides whether a new topic is discussed on a thread or not at the end of each time period through thematic event detection methods.

Chen et al. (2009) proposed a noise-filtered model to extract bursty topics from web forums using terms and participations of users. The study was a thematic plus temporal method which characterized and ranked the terms depending on its frequency of sequential occurrence over time. Another study Chen et al. (2012) also came up with a thematic and temporal method to identify hot topics in web forums. The study defined post of users as document and a topic as a cluster of similar documents with the same semantic description. Each post was defined by contextual features like posts content, timestamp of the post, time segment from the time of first reply to final reply. Both local and global features in terms of term frequency (TF) and global inverse document frequency (GIDF) were considered to find out hot topics in web forum. Single-pass incremental clustering algorithm with a threshold parameter is used for training phase of post content in this study.

A plethora of studies is also present where the researchers have used thematic event detection methods. Periera Nunes et al. (2014) proposed a topic extraction process which facilitated university students and faculties to search and recommend important and relevant topics to be discussed in the online web forum. The study used thematic tools to do named entity recognition and topic extraction. The thematic analysis was followed by a statistical method, calculating the TF-IDF score to select and rank the most representative terms from the users posts of the forum thread. The study extracted the top-ranked topic in the online forum as a final outcome. Another similar study by Devi and Bhaskaran (2015) proposed a method to detect hot spots in online web forum using enhanced *k*-means (*E*-*K*-means) and aging theory. The study defined hot spots as forums that has bulk of thread, posts and discussions and appears quite frequently over a period of time.

Table 5 shows the list of event detection techniques used for web forums categorized under S, T, TH and NS.

Table 5 List of event detection techniques for web forums

Type of streams	Dimension				Techniques used
	S	T	TH	NS	
Web forums				✔	Nonnegative matrix factorization
				✔	Markov logic network
			✔		SVM classifier
			✔		TF-IDF
			✔		UF-ITUF
		✔	✔		Noise-filtered model
		✔	✔		TF-GIDF
			✔		Single-pass incremental clustering
		✔	✔		E-K-means
		✔	✔		Aging theory

2.2.3 Open problems

- *Exploiting web forums for detecting general, trending or emerging events:* Most of the existing studies in web forums considered detecting hot or bursty topics using the web forum content. More studies should be performed on web forum to detect all general topics, trending as well as hot topics by using the content or the user participation as the deciding parameter. The users should also receive a choice of topics to select from and to join in depending on their interest level in the given topics.
- *Improvement in the language pattern by improvising on the semantic features:* The language pattern of web forums is imprecise, casual and terse which makes topic detection a challenging task. So, apart from the existing semantical and contextual features used by researchers in the above studies like posts content, timestamp of the post and the time span of the post, other semantic features such as misspellings, Internet slangs, etc., should also be considered to discover further language patterns in web content of forums.
- *Real-time topic detection in web forums:* Most of the event detection study in web forums implements the forum content in offline mode. Live streaming of web forum posts should be considered to detect real-time events through dynamic addition of new topics. This ensures more topic coverage for the web forum users. For example, in a university framework which includes a student-faculty web forum, the faculties might wish to add new topics dynamically to the forum to enhance more discussions among students. This will further ease out in understanding the topics better through forum than in a classroom environment.
- *Out of context and deviated discussion content from main topic/thread line*: The subject of a thread is not explicitly understandable and is implicit in the content of the discussions for the thread. For example, in a series of threads discussion of T20 World Cup Cricket, a thread title "I think Rohit Sharma should be substituted for!" and with replies, "Yeah, he played worse than last match" might appear. Though this thread talks on the T20 World Cup Cricket event, thread title shows no apparent connection with the event. So, though the threads get selected for event detection purpose, most of the discussions data do not give any output to the researcher.

2.3 Blogs

2.3.1 Description

A Web site that displays the postings of one or more individuals in reverse chronological order and also contains links to comments on some specific postings is defined as blog (Agarwal and Liu 2008). Each entry is called as blog posts, and the sites are called blog sites. Blogs, also termed as weblogs, is a prevalent type of media on Internet (Gill 2005). Many researches are carried out considering blogs as a test bed to prove research problems and algorithms (Kumar et al. 2005; Tseng et al. 2005). The interlinking structure of blogs gives rise to small local communities and can be used to find friendship relationships and proximity of locations (Kumar et al. 2004). Figure 5 illustrates a blog post of blog site of *buzz blogger*. It also consists of specific threads of blogs which are more trendy options for creating new blogs and posting comments for the already existing blogs.

2.3.2 Event detection techniques

The dynamic user-generated content of blogs makes it a very good platform for event detection. A study proposed a novel event detection algorithm by making use of temporally annotated thematic space which tracks the change of semantics of words over time (Jurgens and Stevens 2009). A semantic space model is an automated method of building distributed word representation. The algorithm named temporal random indexing was presented by them that effectively capture changes in semantics or words and phrases over time. These highlighted changes facilitated the researchers to automatically detect new events and to identify associated blog entries through the thematic–temporal event detection method.

A thematic study proposed a novel supervised method for reporting outbreak of disease events as a contribution toward epidemic intelligence (Stewart et al. 2011). They trained a supervised learner and then transferred the learning to classification of blogs reporting disease. This transfer aimed to automatically label the data in the source domain, using a subset of the automatically labeled data as training set to earn a predictive function and then using the knowledge to improve the learning of the target predictive function. The study built a binary, syntactic parse tree-based classifier that is capable of detecting disease reporting sentences in blogs. The study improved its quality of detection by considering three properties of the blog sentences: sentence position, sentence length and sentence semantics.

Most of the studies on blogs focused on offline algorithms which use preaggregated results (Hennig et al. 2014). The study by Hennig et al. (2014) in contrast focused on extracting, analyzing and visualizing the events by aggregating information from many documents of blogs. Given a set of documents and a time interval, the study detected all events generated in this time interval. Each document which was a blog post was tagged by an ID, textual content and its publication timestamp. Extraction of keywords from

CREATE YOUR OWN BLOG!

Blogger offers you instant communication power by letting you post your thoughts to the web whenever the urge strikes. Learn more about it. Or:

WHAT'S UP

Bye-bye, BotW. Hello, BoN. I've retired the Blog of the Week feature. At least for now. I've replaced it with something that will hopefully be both easier to maintain (as you may have noticed, "of the week" came to be somewhat of a misnomer) and more useful/interesting. Introducing: *Blogs of Note* (see the list under Recently Updated). Just a simple, ongoing, irregularly updated list of blogs I've happened to come across and found interesting for one reason or another. This reason need not be substantial. It could be I liked a particular post. It could be the blog seems to have good writing, or good design, or original content or concept, or I just like the name. Anyway, what I'm trying to say is don't read much into the editorial process. Just check 'em out. Or don't.
– Ev. [2/28/2001 05:47:04 PM] +

Part Two of WriteTheWeb: The state of the blog is up.
– Ev. [2/28/2001 04:59:00 PM] +

San Francisco Chronicle: Blogging On: Web loggers bare their souls — and reading lists — to the Internet. Good article on Blogging on the cover of today's Datebook section. Cool. (BTW, the future of the company is much less uncertain than when I talked to that reporter. Things are looking good. Details coming soon.)
– Ev. [2/28/2001 12:45:27 PM] +

WriteTheWeb: The state of the blog. Part one of a three-part email interview I did recently on Blogger's (and blog's) past, present, and future.
– Ev. [2/26/2001 02:57:51 PM] +

The stickers are now swimming up stream to you like little salmon. Thanks a ton to Josh, Randi, Tara, George, Eddie, and Sondra for making it happen. You guys rock.

Fig. 5 Illustration of blogs (*source:* www.buzz.blogger.com)

Table 6 List of event detection techniques in blogs

Type of streams	Dimension				Techniques used
	S	T	TH	NS	
Blogs		✔	✔		Temporal random indexing
			✔		Supervised syntactic parse tree-based classifier
		✔	✔		TF-IDF and named entity recognition (NER)

document was done by TF-IDF and named entity recognition (NER). Identification of bursty intervals was done from the keyword distribution over time intervals. Finally, for creation of events, the study considered cluster of keywords with overlapping bursty intervals through two metrics—co-occurrences and conditional co-occurrences of keywords (Sayyadi et al. 2009). The corresponding formulas for the metric are given as:

$$\text{co-occurrence}\,(k_1, k_2) = \frac{|\{d\epsilon D|k_1 \in d.text \wedge k_2 \in d.text\}|}{|D|}$$

$$\text{conditional\_co-occurrence}\,(k_1|k_2) = \frac{|\{d\epsilon D|k_1 \in d.text \wedge k_2 \in d.text\}|}{|\{d\epsilon D|k_2 \in d.text\}|}$$

where k_1 and k_2 are the two keywords, d is a document, D is the set of documents (corpus), and $|D|$ is the total number of documents.

Table 6 shows all the event detection techniques used by the studies categorized under TH, T, S and NS.

2.3.3 Open problems

- *Required techniques to analyze semantic shift:* The temporal random indexing technique is capable of detecting synonymous event names by identifying words with similar shifts and similar neighbors. But, it fails to analyze the semantic shifts of word index vectors over a period of time. Though techniques like cosine similarity, time-series analysis are well suited for analyzing this shift, it could leave some events undetected as they do not incorporate full information available. So, a proper technique/methods are required to analyze the thematic shift overtime.
- *Reducing the size of semantic slice:* A semantic slice for a word is the chronological ascending order

arrangement of semantics vector of a word in a specific time period which in turn detects event. The size of the semantic slice needs to be narrowed in order to have reduced semantic space. Researchers need to brief the semantic slice without interrupting representation of semantics needed for event detection.

- *Real-time event detection in blogs:* Existing techniques are capable of detecting events through observing the changes of semantic slice at monthly granularity level. But, real-time event detection needs to be operated with much more information into consideration. A relationship should be developed between the corpus, the duration of semantic slice and the types of events detected in order to analyze event detection deeper into the finer scale.
- *Automatic labeling of blogs:* Semi-supervised learning consists of small number of labeled trained data along with large amount of unlabeled data. Automatic labeling or tagging should be studied on a smaller volume of trained dataset for detecting events in blogs. This automatic tagging will improve the process of detecting detailed event and the truth about the event which can be performed by the machine learning algorithm.
- *Event detection using metadata information in blogs:* The metainformation associated with blog entries such as category of the blog, location/origin can be used to improve on the quality of the generated events. Also, events can be augmented with more details on the groups of people associated with the event or the relevant locations with the help of the metainformation. Further research is required to find out how these informative blog entries can be used to find out whether they are discussing the same event.
- *Considering blogs with different timestamps but the same event keywords:* Research is required to frame stories from the blog entries, identifying events having similar keyword but occurring in different time frames. For example, an event of product launch starting from the information getting leaked, followed by the official launch, and then product reaching public can be followed as the story of the same event but occurring at different time frames. This will help the users to have a complete knowledge of the demand, reputation and opinion of customers about the product which is trending in different points of the timeline.
- *Quantification of trust in blogs to determine influencers for event detection:* Influential bloggers are many, but influential blog sites are few (Agarwal et al. 2008). Researchers face challenges to find both influential bloggers and the sites to get facilitated in event detection. The actual challenge lies in quantifying the blogger and the blog post's trust. Several blog sites also allow public to create and edit content, compromising the truthfulness of the original content. Therefore, trust in blogs should be quantified by the researchers through new techniques so that the influence of bloggers as well as the blog sites can be determined before inferring on event detection based on their blogs.

2.4 Email

2.4.1 Description

Communications through email between people as shown in Fig. 6 can be represented graphically with edges representing email communications and vertices showing the email accounts resulting in a communications network (Wan et al. 2009). Huge number of communications through email can be considered as a continuous stream of data. Few researchers have used this stream of data to detect events (Wan et al. 2009).

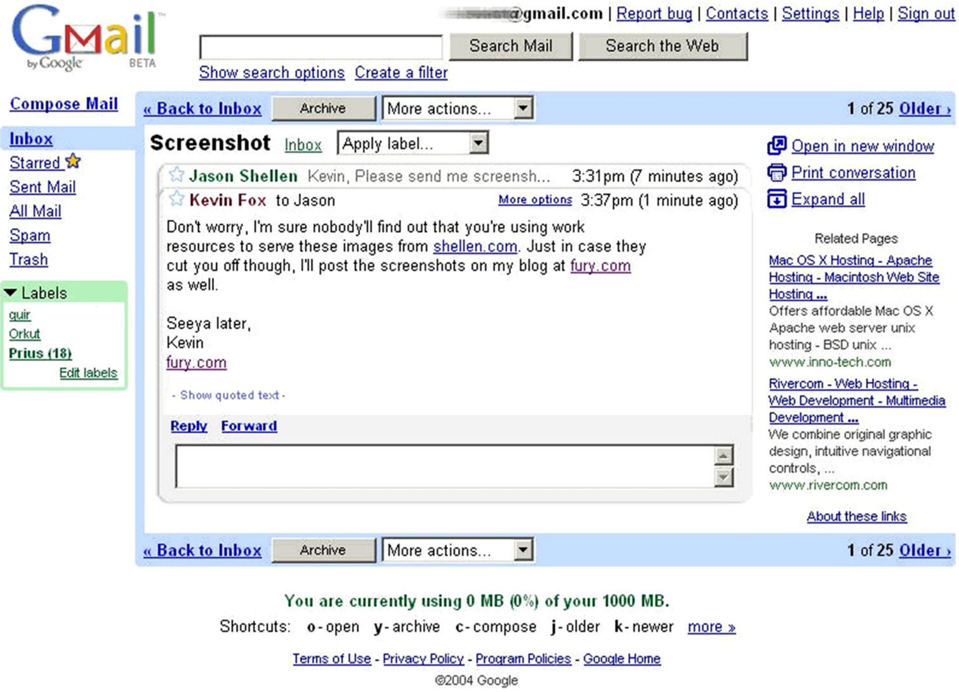
2.4.2 Event detection techniques

A study by Wan et al. (2009) used the linkages in an email network to identify abnormal patterns in email communication during real-world events. The detection was done based on link-based detection where clustering of vertices was done to find out similar communication patterns so that the deviations in patterns between each vertex's personal profile and its cluster profile facilitate in detecting events.

As email communications are temporal in nature, event detection techniques on time-series data were used. For personal profile deviation, temporal methods like hidden Markov model (Ihler et al. 2006), change point detection (Guralnik and Srivastava 1999) and scan statistic can be used.

A temporal, network structural and thematic dimensions were considered in few studies on emails. Zhao and Mitra (2007) used three steps to extract events from social streams of data from Enron email dataset—the text-based clustering, temporal segmentation and graph cuts of OSNs. This study proposed a multidimensional visualization tool which visualizes the relation between events along these three different dimensions. A study on automatic detection of occurrence of events and its contextual information (i.e., location, temporal information, participation of individuals) through email communication was done in Wasi et al. 2011. The study defined an email as event having all the contextual information attached, else the email is considered as a nonevent. Finite state automata (FSA) were used to extract the contextual information of the event. This study has extended part-of-speech (POS) tags to show the transaction within states.

Fig. 6 Illustration of email (*source:* www.fury.com)



Jason Shellen Kevin, Please send me screensh... 3:31pm (7 minutes ago)

Kevin Fox to Jason 3:37pm (1 minute ago)

Don't worry, I'm sure nobody'll find out that you're using work resources to serve these images from shellen.com. Just in case they cut you off though, I'll post the screenshots on my blog at fury.com as well.

Seeya later,
Kevin
fury.com

Related Pages

Mac OS X Hosting - Apache Hosting - Macintosh Web Site Hosting ...
Offers affordable Mac OS X Apache web server unix hosting - BSD unix ...
www.inno-tech.com

Rivercom - Web Hosting - Web Development - Multimedia Development ...
We combine original graphic design, intuitive navigational controls, ...
www.rivercom.com

You are currently using 0 MB (0%) of your 1000 MB.

Table 7 List of event detection techniques used for emails

Type of streams	Dimension				Techniques used
	S	T	TH	NS	
Emails			✔	✔	Link-based clustering
		✔	✔		Hidden Markov model
		✔	✔		Change point detection
		✔	✔		Scan statistic
		✔	✔	✔	Multidimensional technique
	✔	✔		✔	Automatic detection technique using finite state automata

Table 7 shows the categorization of event detection techniques used for emails in different dimensions, viz. TH, S, T and NS.

2.4.3 Open problems

- *Selection of features set in emails:* Selection of feature set is done in emails for differentiating between vertices representing email profile accounts and capturing events. The constraints in doing these are the email streaming data and the large graphs used to represent individual elements in this stream. The other factors which should be considered for selection of features are: Features should exhibit people's communications from different viewpoints, and variations in features should be related to different events of interest. Hence, methods/techniques should be available to assist in selecting features before approaching for the event detection task.
- *Handling emails having two or more event mentions:* Existing studies on emails normally dealt with emails forwarded during a specific event (Wasi et al. 2011; Zhao and Mitra 2007). Event extraction and detection from emails will be more dynamic and robust if techniques are there which are capable of detecting two or more events at the same time from the email content. For example, suppose the subject line of email communication says "Rolling out of Business Continuity Plan (BCP) implementation in the organization: Re." It means that the strategic management intends to intimate all the employees regarding a disaster and all business processes should adhere by the BCP policy till the next intimation from management. This email content can contain either an elaborate mention of many situations and events in the past and make the employees aware and equally serious about it or it can be just one-liner intimation. So, in case of the elaborate mention, subject line and body will have different lists

of event mentions. Also, the body content can be used for detecting more than one event. So, instead of detecting a single predefined event, email content can be used to detect any number of events depending on burst of different event keywords.

- *Use of email attachments for detecting events:* Novel techniques/methods should be constructed to extract the attachments, if any, of the emails depending on the high relevance factor of the email content itself. This is for an email having high percentage of relevant keywords for an event and is expected to have a more relevant document for the event sent as an attachment.
- *Data volume and Privacy regulations:* Voluminous nature of continuous stream of email communications and unavailability of the actual email communication content because of the privacy regulations throws a real challenge to the researchers for using the email data content for detecting events.

2.5 Microblogs

2.5.1 Description

Microblogs forms streams of data on Internet through which users can describe their current status, experiences in short posts distributed via instant messages, mobile phones, email or the web (Java et al. 2007). Microblogging platforms like Twitter, Sina Weibo and Jaiku facilitate easy sharing of status messages among users either publicly or within the respective OSN as illustrated in Fig. 7. The real-time features of these microblog platforms are making microblogs a social hot spot for initializing public events (Zhao et al. 2014). For example, 22 events were reported originally in microblogs in 2010 which is about 16% of the whole 138 events (Xie 2011). Another study cited that this percentage increased to 36% in 2011, i.e., 36% of the total hot real-time events are originally reported to microblogs. There are several studies done on event detection in microblogs which are worth reviewing.

2.5.2 Event detection techniques

A popular microblogging service Twitter has gained much reputation in the researcher's community where the microblogs in form of tweets are used to monitor and detect events (Sakaki et al. 2010). This is because a large number of updates happen in Twitter during any social event like elections, disastrous events like riots or natural calamity.

Some existing studies involve spatiotemporal as well as thematic methods for real-time detection of events on microblogs. Real-time detection of events is done by monitoring the tweets. Sakaki et al. (2010) studied a real-time detection of earthquake event by developing a classifier applying semantic analysis on the tweets and modeling a spatiotemporal model for event detection. SVM was used as a classifier to classify tweets automatically depending on statistical, keyword and word context features in a tweet. In this study, each user is considered as a sensor and the tweets as the sensor information

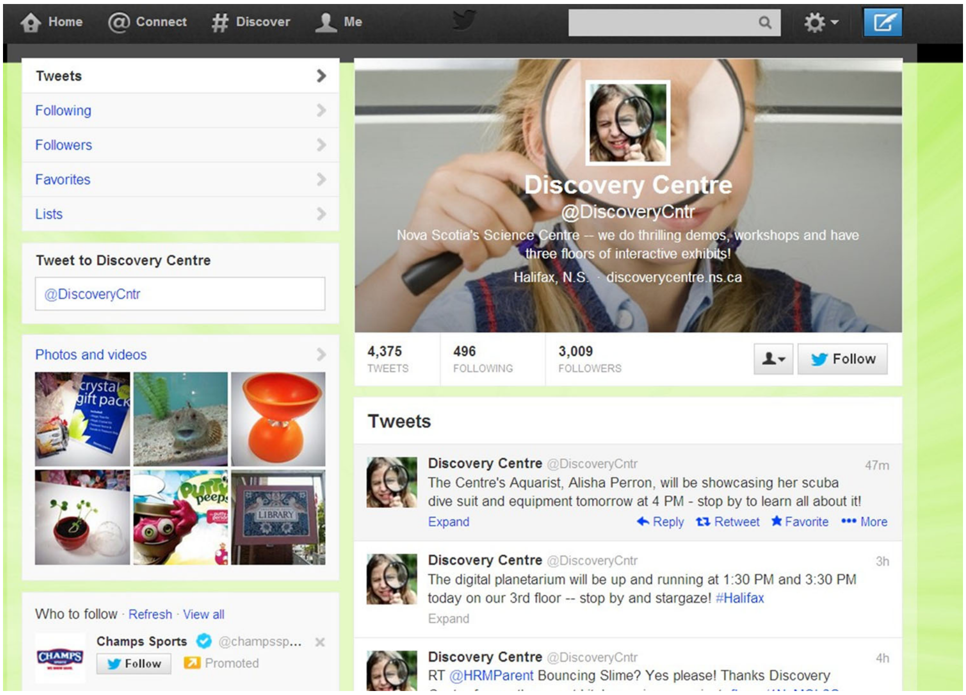


Fig. 7 Illustration of microblogs (*source:* www.rcip-chin.gc.ca)

Nova Scotia's Science Centre -- we do thrilling demos, workshops and have three floors of interactive exhibits!
Halifax, N.S. · discoverycentre.ns.ca

Discovery Centre @DiscoveryCntr 47m
The Centre's Aquarist, Alisha Perron, will be showcasing her scuba dive suit and equipment tomorrow at 4 PM - stop by to learn all about it!
Expand Reply Retweet Favorite More

Discovery Centre @DiscoveryCntr 3h
The digital planetarium will be up and running at 1:30 PM and 3:30 PM today on our 3rd floor -- stop by and stargaze! #Halifax
Expand

Discovery Centre @DiscoveryCntr 4h
RT @HRMParent Bouncing Slime? Yes please! Thanks Discovery

accompanied by timestamp and the location of the user. Event detection was done using the probability density function of the exponential distribution which deals with time-series data. Bayesian filters like Kalman and particle filters are used for location estimation of the tweets. Another pilot study in (Petrovic et al. 2010) dealt with traditional First Story Detection (FSD) in the microblog streaming data setting. The study took a constant time to process each new document along with a constant space which is achieved by modified version of locality-sensitive hashing (LSH) called streaming FSD system.

Automatic online detection of event can also be categorized as a big data task which requires large-scale and intensive real-time stream processing. This was exploited by McCreadie et al. (2013), who proposed an automatic distributed real-time event detection from large volume of tweets. The proposed method can also scale to any volume of input stream without causing any degradation in the performance. Lexical key processing was used to distribute the computational cost of a single document over storm-distributed stream processing platform (McCreadie et al. 2013) instead of categorizing the document stream itself. This was a pilot study done on Twitter firehose where the system can scale up to the entire firehose.

A study by Zhao et al. (2014) classified the event detection in microblogs into three categories—*specific event detection* (Sakaki et al. 2010; Huang and Iwaihara 2011) which pivots around some disaster or some special keywords, *specific person-related event detection* (Popescu and Pennacchiotti 2010; Popescu et al. 2011) which were detected through specific person names or celebrity names and *general events* (Long et al. 2011) which are detected through presence of hot keywords in microblogs. The study presented a framework for event detection from microblog messages containing three modules: *microblog crawler and filtering*, *microblog event detection* and *event prediction.* Microblog platform APIs were used for crawling. Filtering was done by judging the quality of the microblogs based on the user behavior and the content quality based on few metrics like posting date, part of speech and originality. Event detection was done by an object-oriented approach where each evolutional event was modeled by static properties like identifier or attribute descriptor of event and dynamic properties like a spatiotemporal model.

The thematic and temporal features of microblog posts were studied from the Sina Weibo microblog platform in (Li et al. 2014) to detect online bursty events which had been drawing quite a lot of public attention. The study used incremental temporal topic model to track the topic of events drifting over time. Xie et al. (2013) used dimension reduction to detect bursty co-occurrences between keywords through an optimization problem. Another study on bursty event detection by Li et al. (2012a, b) constructed semantically meaningful segments by segmenting the tweets into nonoverlapping segments, each one of which formed as event segments. The dynamic and temporal feature of event evolution was not considered in this study. Weng and Lee (2011) detected event based only on temporal information of the events. The study used wavelet transformation to fit the temporal information of each word. Modularity-based graph partitioning algorithm was used to form events. The document contents of different topics or events, the temporal distribution of the content and the network structure formed through the dynamic interactions of the OSN users were also considered by (Aggarwal and Subbian 2012) while studying the event detection and clustering challenges in social streams. Tweets of Twitter, chat interaction streams over email and chat platforms and the posts on the OSN walls are considered to be a part of social-stream data. Their study proposed novel supervised and unsupervised algorithms for event detection. Both thematic and network structure information were used to create clusters of data streams of events. The study clearly depicts that event detection in social streams based on thematic and network structure information outperforms the existing only thematic-based state-of-the-art studies.

A study defining event by incorporating a composition of multiple event elements over content, time, location and network structure was done by Zhou and Chen (2014). According to the study, a complete view about a situation is needed for right decision during crisis. Multiple events can be detected simultaneously and correctly by analyzing an overall view of the situation depicted through microblogs. To solve this, three requirements should be prioritized—*firstly*, a robust data representation model which captures the contexts of content, time, location and social information of the microblogs; *secondly,* an advanced technique to handle uncertainty of data; and *thirdly,* a set of efficient query processing techniques to enhance the understanding of social messages. Accordingly, the study proposed a new graphical model called location–time-constrained topic to capture social microblogs over content, time and location. They also proposed a complementary measure to find the similarity between messages with content, time, location and the links in the network structure. Finally, they implemented a similarity join over microblog of Twitter and design hash-based index scheme to improve the efficient of event detection.

Table 8 lists the event detection techniques used in microblogs and are categorized under TH, T, S, and NS categories.

Table 8 List of event detection methods used for microblogs

Type of streams	Dimension				Techniques used
	S	T	TH	NS	
Microblogs			✓		SVM
		✓			Probability density function of exponential distribution
	✓				Kalman and particle filter
			✓		Lexical key processing
	✓	✓			Locality-sensitive hashing
		✓			Probabilistic temporal model
		✓			Incremental temporal model
		✓			Object-oriented approach
	✓	✓	✓	✓	Location–time-constrained topic
		✓			Wavelet transformation
			✓		Hash-based indexed scheme

2.5.3 Open problems

- *Dynamic update of topics/events without any predetermined mention of event:* Most of the studies in the microblog platform consider the number of topics to be predetermined. For example, researchers opt for a conscious detection of events during a specific timeline where they are aware of the events happened. In such cases, a query-based or keyword-based search is used to support the event detection. But, in real-time scenario, the events which are going to occur are not always known to the users. So, further studies are required to develop techniques where the number of topics will be updated dynamically along the timeline without any mention of the predetermined events.
- *Efficient platform and techniques to deal with the incremental data:* Microblog streaming data require robust framework or platform accompanied by efficient techniques to scale up whenever there is a need. Hence, further work should be done by the researchers on the model like Spark (Apache Spark 2016), which is a distributed in-memory computing platform to handle larger datasets in microblog platform for event detection.
- *Dearth of techniques coupling textual, spatial and temporal along with social/network structure:* There is no model or structure that exists for events in microblog platform. This leads to different definitions of events made by the researchers in context to their research problem. Also, study (Zhou and Chen 2014) portrays that social or network structure properties of microblog platforms should be exploited more to increase the granularity of event detection. Most of the event detection in microblog platform happened with textual content. But, to have a more accurate detection, textual along with temporal, spatial and social/network structural properties should be considered.

3 List of tools used for event detection in OSN platforms

A detailed review of different event detection techniques in OSN for various types of streams of data was done in the previous sections. In this section, a detailed review of the tools available for event detection for these various data streams is done.

List of existing tools which facilitates in detecting and analyzing events is shown in Table 9. These are end-to-end tools which—(1) apply different natural language processing techniques on the raw data from the OSN platform for filtering; (2) perform different analyses, for, e.g., sentimental analysis, finding emerging patterns, running meaningful queries, trend analysis, on the filtered data; and (3) incorporate techniques for meaningful visualization of the data. Most of the tools have incorporated spatio, temporal and thematic analysis of data with very less contribution toward the social/network structure. All the surveyed tools were used to find the pattern of real-time event or bursty events in OSN. In doing so, each study tried to improve on the complexity of the algorithm employed to ensure consistent performance of the system in finding the pattern. Most of the tools as shown in Table 9 are using microblogs as input data stream except a few like Blogscope and NodeXL, which are effective in analyzing blogs, web forums or newswires.

A discussion about the tool, along with features, and its purpose highlighting the extent to which the tool is capable of participating in the event detection and analysis tasks from different social media data streams are as follows:

Table 9 List of event detection tools under several OSN platforms

Tools (Year)	Types of streams	Feature	Types of events	Types of event dimensions
ReDites (2014)	Microblogs	Real-time event detection tool for information security analysts	NMDE/ EE/ POE	TH, T, S
TopicSketch (2013)	Microblogs	Real-time bursty event detection without any predefined topical keywords	NMDE/ EE/ POE	TH, T
TweetXplorer (2013)	Microblogs	Prominent visualization tool depicting events timeline, significant tweets, users who retweets, location and user pattern display	NMDE	TH, T, S, NS
Social Sensor (2013)	Newswire, blogs, microblogs	Processes microblogs and multimedia data to automatically discover event, influencers, trends and interesting content	NMDE/ EE/ POE	TH, T, S, NS
TEDAS (2012)	Microblogs	Detects and ranks new events from crime and disaster-related events both in offline and in online modes	NMDE	TH, T, S
Twevent (2012)	Microblogs	Bursty event detection with a predefined topic name	NMDE/ EE/ POE	TH, T
Twitcident (2012)	Newswires, blogs, microblogs	Detects events by automatically connecting to emergency broadcasting services	NMDE/ EE/ POE	TH
TweetTracker (2011)	Microblogs	Provide information to the first responders regarding relief measures decision making during HADR	NMDE	TH, T, S
Twitinfo (2011)	Microblogs	Real-time tool which provides a graphical view of subevents and also provides summarization of the event	NMDE/ EE/ POE	TH, T, S
TwitterMonitor (2011)	Microblogs	Real-time visualization tools which detect bursty keywords, trending topics along a specific timeline	EE	TH, T, S
SensePlace2 (2011)	Microblog, blogs, newswire	Geovisual analytics application which provides visually enabled sensemaking on the data	NMDE	TH, T, S
NodeXL (2011)	Newswire, web forums, blogs, microblogs	Aids in visualizing different types of networks and performs different graph metric calculation	NMDE/ EE/ POE	TH, T, S, NS
Twitris (2009)	Newswire, web forum, blogs, microblogs	Performs a brand management for the companies depending on the data available from the customers	NMDE/ EE/ POE	TH, T, S, NS
Trendsmap (2009)	Microblogs	Use archival data to derive the sentiments of the most engaged areas through maps	NMDE/ EE/ POE	TH, T, S
Memetracker (2009)	Newswires, blogs	Uses phrases and quotes to map the daily news cycle to the newswires stories and available blog posts	NMDE/ EE/ POE	TH, T
Ushahidi (2008)	Web forums, email, microblogs	Crowdsourcing-based content management system used by organizations for decision making	NMDE/ EE/ POE	TH, T
Blogscope (2007)	Blogs	Provides competitive intelligence and market trends to the users through spatiotemporal analysis of blogs	POE	TH, T, S

a. *ReDites* (Osborne et al. 2014).
About: ReDites is a tool which aids in real-time event detection, tracking, monitoring and visualization, specially designed for information analysts of security sector. It dealt with large-scale data and tailors it to the security domain. The tool was tried with the dataset of terrorist attack that happened on September 2013.
Features: Event processing comprises of four steps. Firstly, new events are detected from the first tweet itself. Next, it tracks the event, searching for new posts relating to the single tweet and keeping an update on the incremental study of tweets. Next, the events are organized and categorized for the security domain, and geolocation is performed and detected for the

sentiment that evolves around that event. Finally, the produced stream is visualized, summarized, categorized for the information security analysts.
Purpose: Real-time event detection tool which supports large-scale incremental microblog streaming data.

b. *TopicSketch* (Xie et al. 2013).
About: TopicSketch leverages Twitter for automated real-time bursty topic detection as soon as it occurs. The tool was tested on a stream containing over 30 million tweets and demonstrated the effectiveness and efficiency of their method.
Features: The tool provides temporally ordered subevents that are more descriptive in nature and can also detect events with bursts over shorter duration of time. Its contribution can be divided into three stages—firstly, it performs a data sketch which does a calculation on the total number of tweets, the occurrence of each word and the occurrence of each word pair which provides an early indication to the popularity of a tweet. Depending on this, a topic model was developed to infer on bursty topics whose dynamism overtime is calculated through data sketch. In the second stage, tool performs a hashing-based dimension reduction technique using hashing to achieve scalability and maintain quality of the events with proved error bounds. The tool is scalable to an extent of handling 300 million tweets per day which is close to the data which gets generated on Twitter platform in a day.
Purpose: For dynamic update of topics/events without any predetermined event mention from a very large-scale data.

c. *TweetXplorer* (Morstatter et al. 2013).
About: TweetXplorer is an analyst's tool to gain knowledge on social big data through effective visualization techniques. The dataset used was Hurricane Sandy through which they exhibit the working of the system.
Features: This tool explored the Twitter data by following the data lifecycle phases which includes "Plan and prepare," "Collect and process," "Analyze and Summarize," "Represent and Communicate" and "Implement and Manage." TweetTracker (see h) support is for the first two phases. The main functionalities of the tool are in creating meaningful queries, discovering interesting time periods, representing important tweets and users depending on retweet facility, communicating salient locations and discovering user patterns. D3 visualization toolkit is used with the force-directed layout method for the visualization process.
Purpose: To convey information of any disaster event visually so that necessary actions can be taken during crisis.

d. *Social Sensor* (Papadopoulos et al. 2015).
About: Social Sensor is a project which collects processes and aggregates large streams of social media and multimedia data. It outputs the trends, event influencers and interesting media content to the users by analyzing these huge streaming data. It uses Twitter, Facebook, Youtube, Instagram and other social media platforms to collect data.
Features: It works on real-time data and automatically discovers important trends and cluster of events. It also supports a verification process while handling text, images, audio and video. It organizes the content by location, time, sentiment and influence.
Purpose: To aggregate streaming data from various large-scale data sources.

e. *TEDAS (Event detection and analysis system)* (Li et al. 2012a, b).
About: TEDAS detects and analyzes events through tweets from the Twitter platform. The dataset used by the researchers for the tool is crime and disaster-related events, for example car accidents and earthquake.
Features: The three important functionalities are detecting new events, ranking the events w.r.t. their importance and generating temporal and spatial patterns of the event. It works on both offline and online computing mode. It detects new events maintaining a rule-based approach. The classifiers, the metainformation extractor and text search engines are used for extracting informative tweets and extracting location and temporal details from the same. The visualization of the event is done on the basis of a given keyword and the timeline within which the event visualization is expected. The implementation of this tool was done based on Java, PHP with the backend support of MySQL, Lucene, Twitter API and Google Maps API.
Purpose: Event detection through Twitter. This tool uses metadata information in case of blogs.

f. *Twevent* (Li et al. 2012a, b).
About: Twevent is an online segment-based bursty event detection tool for tweets with a predefined topic name. The event detection was done using Twevent based on the 4.3 million tweets published by Singapore-based users in June 2010.
Features: The tool detects the bursty tweet segments as event segments and then performs a clustering using their content similarity and frequency distribution. The nonoverlapping segments formed from each tweet are semantically meaningful information units. Every bursty segment is identified within a fixed time window based on frequency patterns. After finding out the candidate events, Wikipedia (Wikipedia, 2016) is used to identify the most realistic events to report the final identified events.

Purpose: To detect bursty event through bursty tweet segments as event segments.

g. *Twitcident* (Abel et al. 2012).

About: Twitcident is a web-based system which provides a framework for automatically filtering relevant information from social media streams, searching for events and analyzing information regarding this real-world incidents or crisis. This tool automatically connects to the emergency broadcasting services and start tracking and filtering the new incident or event from the social media streams.

Features: The incident detection module senses the incidents from the broadcast emergency services. When this new thread of incident is reported by Twitcident core framework, the tool starts collecting and aggregating related messages from the web and Twitter. These messages are further processed by the semantic enrichment module which features named entity recognition, classification of messages, linkages between messages to external web resources and the extraction of metadata. Additionally, users are also provided with a search option through which they could further receive filtered messages according to their need. The tool also provides graphical visualization of the evolution of the incident overtime or the geographical impact on the area of an incident.

Purpose: Enhance the user involvement through finer level of visualization (as required by the users).

h. *TweetTracker* (Kumar et al. 2011).

About: TweetTracker is an application designed to facilitate the Humanitarian Aid and Disaster Relief (HADR) organizations to track, analyze and monitor the disaster-related tweets from the Twitter platform. The aim to design this tool is to provide first responders achieve proper whereabouts about the disaster situation to decide on the relief measures. It has used the tweets of Cholera crisis happened in Haiti to validate the functioning of the tool.

Features: The disaster relief operations require real-time monitoring of the tweets within a very short span of time from the time of the disaster. This tool helps in real-time monitoring by analyzing the tweets from temporal, geospatial and topical perspectives. Twitter streaming API is used to collect tweets, and filtering of informative tweets is done based on specific keywords, hashtags and geolocation of the tweets. Keywords are used to find the trends through the keyword trending engine and are expressed through tag clouds. The map is used for showing the geolocated tweets. This tool also works on the multilingual tweets through Google Translate to enhance understanding of the tweets.

Purpose: Real-time detection of events with predefined event mentions from users.

i. *Twitinfo* (Marcus et al. 2011).

About: Twitinfo is a real-time tool for visualizing and summarizing events on Twitter. It allows to browse huge amount of tweets about different events like disaster, politics and sport and use a timeline-based display to show the peaks of high tweets activity. According to expert opinion, this tool is appropriate to monitor long-running events and also to identify the eyewitnesses.

Features: The tool can extract the tweets matching the keywords present in the query and output a graphical view of the timeline of the subevent through peaks of tweets reaching a certain number. It also highlights important terms and messages concerning the subevent, provides zoom in facility to the users, does a sentiment analysis displaying the event-related user's sentiment's and also the geographical distribution of tweets. It also provides the links between tweets in a cluster if it is portraying any relevance.

Purpose: Real-time analysis of tweets for event detection.

j. *TwitterMonitor* (Mathioudakis and Koudas. 2010).

About: TwitterMonitor is a visualization tool which shows the trend of an event in real time on Twitter along a specific timeline and its analysis.

Features: The tool starts with detecting the bursty keywords in the streaming data of tweets which is treated as a starting point of detecting a new event. QueueBurst and GroupBurst algorithms were developed to find out bursty keywords in real time. The most correlated keywords were found through context extraction algorithms like principle component analysis (PCA), singular value decomposition (SVD) using latent semantic analysis which finds out the most correlated keywords within the event. Grapevine's Entity Extractor's algorithm is used to find the most frequently maintained entities in the trends. It generates a chart that depicts the evolution of popularity of event overtime along with the geographical origins of tweets. The trending topics of different events extracted are also indexed according to the volume or a recency score or the combined score of the both.

Purpose: To index trending topics depending on the volume and newness of tweets, along with its timeline-based visualization.

k. *SensePlace2* (MacEachren et al. 2011).

About: SensePlace2 is a geovisual analytics application that collects tweets attributed by place–time attribute information from all the Twitter users and supports crisis management during disaster events through visually enabled sense making from the available information.

Features: The tool uses a crawler which via Twitter API collects tweets of interest using keyword and hashtags. Tweets and auxiliary metadata are stored in JSON format. These data are parsed and stored in a PostgresQL database. Different distributed applications that need to analyze tweets for named entities such as location, entities and hashtags then work on the database and create respective tables. Lastly, a LUCENE text index is generated which supports a full-text retrieval of tweets within geographical region and data range. The visualization supports the analysts to explore, characterize and compare the spatial, temporal and geography associated with the topics and the entities of the tweets.

Purpose: Supports full-text retrieval of tweets within a particular geographical region and data range.

l. *NodeXL* (MarcSmith 2016).

About: NodeXL facilitates in collecting, analyzing and visualizing a variety of networks.

Features: It allows for flexible import and export of different data source files, populated with varied types of data fields like timestamp, geolocation and thematic data. The tool can establish a direct connection to OSN and involve dynamic filtering of network features and powerful vertex grouping. The visualization of graph provides zoom in and scaling facility into the interest areas of network graph and performs different graph metric calculation and flexible layout done by forced directed algorithm.

Purpose: Perform dynamic filtering of network features and powerful vertex grouping.

m. *Twitris* (Purohit and Sheth 2013).

About: Twitris helps in brand management for the companies with an effective sentiment and emotion analysis done on the brand as well as a comparative study between the competitors regarding their performance, strategies and products. It also shows several trending topics or breaking news trending at that point of time.

Features: The tool analyses the spatial, temporal and thematic aspects of the tweets and extracts the event descriptors. The data semantics is captured in three contexts: *internal context* (images, videos, other web content directly related to the tweet, other tweets containing the same event descriptors and semantically annotated entities mentioned in the tweet), *external context* (external sources like Google news, Wikipedia articles and other news media sources like BBC) and *mined internal context* (context obtained by mining the internal context). The meaning of the event descriptors is understood through the use of deep semantics using automatically created domain models. Shallow semantics were used for knowledge discovery and representation. Finally, the processed social data are exposed to the public domain utilizing a semantic similarity with other external web resources like news, articles, images and videos.

Purpose: An improved tool which works on relation detection and co-reference resolution detecting events from various sources and validating events from external media like television.

n. *Trendsmap* (Trendsmap 2015).

About: Trendsmap shows the latest trend in the Twitter from anywhere in the world. It can be used in agile and content market, brand management, crisis management and trend monitoring.

Features: This tool helps in finding historical trends by analyzing historical data as far as back in mid-2009s or using their extensive archive data. It involves sentiment analysis involving the most engaged areas through map. Relevant data are filtered depending on location, time language, etc. It also provides a word cloud for specific countries.

Purpose: Finds trends from historical or archival data of the company to provide word cloud for specific countries.

o. *Memetracker* (Leskovec et al. 2009).

About: Memetracker builds maps of the daily news cycle by analyzing newswires stories and blog posts, ranging from the mass media to personal blogs. The tool used a huge dataset of 90 million newswires and blog posts, collected over a duration of 3 months during the 2008 US Presidential Elections. The selection of US election data to study and analyze events made the tool popular.

Features: The tool helps the users to see how different stories compete for news and blogs coverage every day and how some of the stories remain there for certain period of time and some get wiped off quickly. The tool firstly identifies short distinctive phrases that are present in blogs or newswires and evolve over time. These phrases are considered as “genetic signatures” for different newswires/blogs which experiences a significant mutation overtime. Finally, these phrases from different newswires/blogs are clustered which contains all mutational variants of a single phrase.

Purpose: The tool analyzes the semantic shift of events over a period of time. It identifies stories which are popular, and stories which fades away with time across all newswires and blogs.

p. *Ushahidi* (Ushahidi 2008).

About: Ushahidi uses the concept of crowdsourcing to help the people tackle most challenging situations in this world. It gives the opportunity to the users to upload eyewitness reports during any disaster or crisis which further gets visualized as a map by the Ushahidi software. It is used by many organizations for brand management and decision making.

Features: This platform involves information collection, visualization and interactive mapping. It consists of SwiftRiver platform which enables filtering and verifies real-time data from different data sources like Twitter and RSS. The Crowdmap platform allows the deployment of Ushahidi platform without having to install it in the web server.
Purpose: Real-time topic detection in web forums.

q. *Blogscope* (Bansal et al. 2007).
About: Blogscope tool provides competitive intelligence information and market trends to the users using blogs.
Features: It is an information discovery and text analysis system which includes features like spatial–temporal analysis of blogs. Its functionality involves finding of keyword correlations, geosearch, hot keywords and burst synopsis generation. Events are detected with the occurrence of a burst of hot keywords at a particular time in reply to a query.
Purpose: Burst identification by undergoing real-time event detection of blogs and through analyzing semantic shift over a timeline.

A summarization of the above list of tools is given in Table 9. The tools are categorized according to year of its development, type of input streams, distinguished feature, effective types of events and the type of event dimensions applied for its development.

After comparing the features of the tools, we found that few of the open challenges that were discussed in Sects. 2.1–2.5 have been addressed through the development of these tools. They are as follows:

- The requirements of suitable platform and techniques to deal with the real-time large-scale incremental microblog streaming data are solved by *ReDites*. It is effective for information analysts in security domain but can be checked to generalize for any event over microblog platform.
- Dynamic update of topic/events without the need of any predetermined event mention from a large scale of data can be done through the *TopicSketch*.
- The real-time detection of events with predefined event mentions from the users is done by *TweetTracker*.
- The use of metainformation for event detection is done by *TEDAS* and *Twitcident* for microblogs.
- The enhancement of procedures to involve users more by providing finer level visualization, zooming facility, on demand geographical distribution of tweets is done by both *Twitinfo* and *Twitcident*.
- *Memetracker* is an excellent tool for newswires and blogs which analyzes the semantic shift of events over a period of time. It is capable of detecting stories which gets popular at a certain time and also stories which is there for a short duration of time.
- One of the web forum issues of real-time topic detection gets solved through *Ushahidi* tool which is excellent for content management for organizations.
- *Blogscope* does real-time burst identification in blogs through analyzing semantic shift over a given timeline.

4 Conclusion

OSN platform has become a prime platform for researchers to get relevant information by churning the huge social-stream data available during different kinds of events. The relevant information helps to take appropriate actions by the authorities in case of natural disasters, trending or emerging topics and public opinion-based events.

This paper first described the various dimensions—thematic, temporal, spatial and network oriented, in which the event can be categorized, and then, surveyed event detection techniques for social text stream—newswire, web forms, blogs, emails and microblogs. The open challenges were also discussed for each text stream that will be helpful for the researchers to work on. Toward the end, tools available for event detection were described and few open issues that they solved were discussed.

We tried to focus on event detection techniques and event detection tools separately and then mapped into using open challenges of the former with features of the latter. The insights were useful in the sense that researchers can contribute on the open challenges (as suggested in the event detection techniques) not only in the form of new techniques but also in the form of new tools.

References

Abel F, Hauff C, Houben GJ, Stronkman R, Tao K (2012) Twitcident: fighting fire with information from social web streams. In: Proceedings of the 21st international ACM conference companion on world wide web 305–308. doi: 10.1145/2187980.2188035

Agarwal N, Liu H (2008) Blogosphere: research issues, tools, and applications. ACM SIGKDD Explor Newsl 10(1):18–31. doi:10.1145/1412734.1412737

Agarwal N, Liu H, Tang L, Yu PS (2008) Identifying the influential bloggers in a community. In: Proceedings of the 2008 ACM international conference on web search and data mining (WSDM'08) 207–218. doi:10.1145/1341531.1341559

Aggarwal C, Subbian K (2012) Event detection in social streams. In: Proceedings of the 2012 SIAM international conference on data mining 12:624–635

Aggarwal CC, Zhai C (2012) Mining text data. Springer, Berlin

Ahn D (2006) The stages of event extraction. In: Proceedings of the workshop on annotating and reasoning about time and events. Association for Computational Linguistics 1–8

Allan J, Papka R, Lavrenko V (1998) On-line new event detection and tracking. In: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval 37–45. doi:10.1145/290941.290954

Apache Software Foundation (2016) Apache Spark. http://spark.apache.org/. Accessed 1 Nov 2016

Atefeh F, Khreich W (2015) A survey of techniques for event detection in twitter. Comput Intell 31(1):132–164. doi:10.1111/coin.12017

Balazinska M (2007) Event detection in mobile sensor networks. In: National Science Foundation (NSF) workshop on data management for mobile sensor networks 2007 (MobiSensors)

Bamrah NH, Satpute BS, Patil P (2014) Web forum crawling techniques. Int J Comput Appl 85:17

Bansal N, Koudas N (2007) Blogscope: spatio-temporal Analysis of the blogosphere. In: Proceedings of the 16th international ACM conference on world wide web 1269–1270. doi:10.1145/1242572.1242802

Becker H, Naaman M, Gravano L (2011) Beyond trending topics: real-world event Identification on twitter. Int Conf Web Soc Media 11:438–441

Benson E, Haghighi A, Barzilay R (2011) Event discovery in social media feeds. In: Proceedings of the 49th annual meeting of the association for computational linguistics. Human Language Technologies 1:389–398

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

Chen CC, Chen MC (2008) TSCAN: A novel method for topic summarization and content anatomy. In: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval 579–586. doi:10.1145/1390334.1390433

Chen Y, Yang S, Cheng X (2009) Bursty topics extraction for web forums. In: Proceedings of the eleventh international ACM workshop on Web information and data management 55–58

Chen F, Du J, Qian W, Zhou A (2012) Topic detection over online forum. In: Web information systems and applications IEEE ninth conference (WISA) 235–240

Cheng V, Li CH (2007) Topic detection via participation using Markov logic network. In: Signal-image technologies and internet based system. Third international IEEE conference, 85–91

Chester TLS, Taylor M, Sandhu J, Forsting S, Ellis A, Stirling R, Galanis E (2011) Use of a web forum and an online questionnaire in the detection and investigation of an outbreak. Online journal of public healthinformatics3.1

Cisco VNI (2014) The zettabyte era: trends and analysis. Updated (29/05/2013). http://www.cisco.com/c/en/us/solutions/collateral/serviceprovider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html. Accessed Jan 2016

Cordeiro M (2012) Twitter event detection: combining wavelet analysis and topic inference summarization. In: Doctoral Symposium on Informatics Engineering (DSIE'2012)

Dasigi P, Hovy EH (2014) Modeling newswire events using neural networks for anomaly detection. In: 25th International Conference on Computational Linguistics (COLING 2014) 1414–1422

Deitrick W, Hu W (2013) Mutually enhancing community detection and sentiment analysis on twitter networks. J Data Anal Inf Process 1:19–29

Dereszynski E, Dietterich T (2007) Probabilistic models for anomaly detection in remote sensor data streams. In: Proceedings of the 23rd conference on Uncertainty in Artificial Intelligence (UAI-2007) 75–82

Devi KN, Bhaskaran VM (2015) Online forums hotspot detection and analysis using aging theory. World Academy Sci Eng Technol Int J Comp Electr Autom Control Inf Eng 9(4):913–917

Facebook (2016). www.facebook.com. Accessed December 2015

Fox D, Hightower J, Liao L, Schulz D, Borriello G (2003) Bayesian filtering for location estimation. IEEE Pervasive Comput 3:24–33

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Annals Stat 29:1189–1232

Fung GPC, Yu JX, Yu PS, Lu H (2005) Parameter free bursty events detection in text streams. In: Proceedings of the 31st international conference on Very large data bases 181–192

Gill KE (2005) Blogging, RSS and the information landscape: a look at online news. In: WWW 2005 workshop on the weblogging ecosystem

Gionis A, Indyk P, Motwani R (1999) Similarity search in high dimensions via hashing. In: Very large data bases 1999 Sep 7 (VLDB) 99:518–529

Goller C, Kuchler A (1996) Learning task-dependent distributed representations by Backpropagation through structure. In: Neural Networks. 1996 IEEE International Conference 1:347–352

Grishman R, Westbrook D, Meyers A (2005) NYU's english ACE 2005 system description. In: Proceedings of ACE 2005 Evaluation Workshop, Washington

Gu H, Xie X, Lv Q, Ruan Y, Shang L (2011) Etree: effective and efficient event modeling for real-time online social media networks. In: Web Intelligence and Intelligent Agent Technology (WI-IAT). 2011 IEEE/WIC/ACM International Conference 1:300–307

Guralnik V, Srivastava J (1999) Event detection from time series data. In: Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining 33–42

Halliday M, Hasan R (1976) Cohesion in english. Longman, London

Hardy H, Kanchakouskaya V, Strzalkowski T (2006) Automatic event classification using surface text features. In: Proc. AAAI06 workshop on event extraction and synthesis 36–41

He Q, Chang K, Lim EP (2007) Analyzing feature trajectories for event detection. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in Information Retrieval 207–214

He Q, Chang K, Lim EP, Zhang J (2007) Bursty feature representation for clustering text streams. In: SIAM International Conference of Data Mining 491–496

Hennig P, Berger P, Kurzynski D, Rantzsch H, Meinel C (2014) Efficient event detection for the blogosphere. In: Big Data and Cloud Computing (BdCloud). 2014 IEEE Fourth International Conference 408–415

Hong Y, Zhang J, Ma B, Yao J, Zhou G, Zhu Q (2011) Using cross-entity inference to improve event extraction. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies 1:1127–1136)

Huang J, Iwaihara M (2011) Realtime social sensing of support rate for microblogging. In: 2011 International Springer Conference of Database Systems for Advanced Applications 357–368

Ihler A, Hutchins J, Smyth P (2006) Adaptive event detection with time-varying poisson processes. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 207–216

Jadhav AS, Purohit H, Kapanipathi P, Anantharam P, Ranabahu AH, Nguyen V, Sheth AP (2010) Twitris 2.0: semantically empowered system for understanding perceptions from social data. http://corescholar.libraries.wright.edu/knoesis/252. Accessed Oct 2016

Java A, Song X, Finin T, Tseng B (2007) Why we twitter: understanding microblogging usage and communities. In:

Proceedings of the 9th WebKDD and 1st ACM SNA-KDD 2007 workshop on Web mining and Social Network Analysis 56–65

Ji H, Grishman R (2008) Refining event extraction through cross-document inference. In: Association for Computational Linguistics (ACL) 254–262

Joachims T (1998) Text categorization with support vector machines: learning with many relevant features. In: Proceedings of the Tenth European Conference on Machine Learning 137–142

Jurgens D, Stevens K (2009) Event detection in blogs using temporal random indexing. In: Association for Computational Linguistics Proceedings of the Workshop on Events in Emerging Text Types 9–16

Kaplan AM, Haenlein M (2010) Users of the world, Unite! The challenges and opportunities of social media. Business Horizons 53(1):59–68

Kastner I, Monz C (2009) Automatic single-document key fact extraction from newswire articles. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics 415–423

Keogh EJ (2002) Exact indexing of dynamic time warping. Knowl Inf Syst 7(3):358–386

Li C, Sun A, Datta A (2012) Twevent: segment-based event detection from tweets. In: Proceedings of the 21st ACM international conference on information and knowledge management 155–164

CNN Library (2015) Mumbai terror attacks fast facts. http://edition.cnn.com/2013/09/18/world/asia/mumbai-terror-attacks/. Accessed Jan 2016

Kerman MC et al. (2009) Event detection challenges, methods, and applications in natural and artificial systems. In: Proceedings of 14th International Command and Control Research and Technology Symposium: “C2 and Agility”

Menon R. Gulati A (2010) Spatial—Temporal random indexing for event detection in newswire data, http://ankushgulati.weebly.com/uploads/6/0/3/6/6036818/final_report.pdf, Accessed Jan 2016

Khreich W, Granger E, Miri A, Sabourin R (2012) A survey of techniques for incremental learning of HMM parameters. Inf Sci 197:105–130

Kleinberg J (2006) Data stream management: processing high-speed data streams. Chapter temporal dynamics of on-line information streams. Springer, Berlin

Kumar R, Novak J, Raghavan P, Tomkins A (2004) Structure and evolution of blogspace. Commun ACM 47(12):35–39

Kumar R, Novak J, Raghavan P, Tomkins A (2005) On the bursty evolution of blogspace. World Wide Web 8(2):159–178

Kumar S, Barbier G, Abbasi MA, Liu H (2011) TweetTracker: an analysis tool for humanitarian and disaster relief. In: International Conference on Web and Social Media (*ICWSM)*2011 Jul 5

Kumaran G, Allan J (2004) Text classification and named entities for new event detection. In: Proceedings of the 27th Annual international ACM SIGIR conference on Research and development in information retrieval 297–304

Lam W, Meng HML, Wong KL, Yen JCH (2001) Using contextual analysis for news event detection. Int J Intell Syst 16(4):525–546

Leskovec J, Backstrom L, Kleinberg J (2009) Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining 497–506

Li R, Lei KH, Khadiwala R, Chang KCC (2012) Tedas: a twitter-based event detection and analysis system. In: Data engineering (icde), 2012 IEEE 28th international conference 1273–1276

Li Q, Ji H, Huang L (2013) Joint event extraction via structured prediction with global features. Assoc Comput Linguist 1:73–82

Li J, Tai Z, Zhang R, Yu W, Liu L (2014) Online bursty event detection from microblog. In: Utility and Cloud Computing (UCC) 2014 IEEE/ACM 7th International Conference 865–870

Liao S, Grishman R (2010) Using document level cross-event inference to improve event extraction. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics 789–797

MacEachren AM, Jaiswal A, Robinson AC, Pezanowski S, Savelyev A, Mitra P, Blanford J (2011) Senseplace2: geotwitter analytics support for situational awareness. In: Visual Analytics Science and Technology (VAST), 2011 IEEE Conference 181–190

Madani A, Boussaid O, Zegour DE (2014) What’s happening: a survey of tweets event detection. In: Proceedings of the 3rd International Conference on Communication, Computation, Networks and Technologies INNOV2014 16–22

MarcSmith (2016) NodeXL: Network overview, discovery and exploration of excel. http://nodexl.codeplex.com/, Accessed Jan 2016

Marcus A, Bernstein MS, Badar O, Karger DR, Madden S, Miller RC (2011) Twitinfo: aggregating and visualizing microblogs for event exploration. In: Proceedings of the ACM SIGCHI conference on Human factors in computing systems 227–236

Margineantu D, Wong WK, Dash D (2010) Machine learning algorithms for event detection: A special issue of Machine Learning Journal. Springer 79: 257–259

Maslennikov M, Chua TS (2007) June) A Multi-Resolution Framework for Information Extraction from Free Text. Annual Meeting-Association for Computational Linguistics 45(1):592

Massoudi K, Tsagkias M, De Rijke M, Weerkamp W (2011) Incorporating query expansion and quality indicators in searching microblog posts. In: European Springer Conference on Advances in information retrieval 362–367

Mathioudakis M, Koudas N (2010) Twittermonitor: trend detection over the twitter stream. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data 1155–1158

McCreadie R., Macdonald C, Ounis I, Osborne M, Petrovic S (2013) Scalable distributed event detection for twitter. In: Big Data, 2013 IEEE International Conference 543–549

Metzler D, Bernstein Y, Croft WB, Moffat A, Zobel J (2005) The recap system for identifying information flow. In: Proceedings of the 28th Annual International ACM Special Interest Group of Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval 678–678

Metzler D, Cai C, Hovy E (2012) Structured event retrieval over microblog archives. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 646–655

Miwa M, Thompson P, Korkontzelos I, Ananiadou S (2014). Comparable study of event extraction in newswire and biomedical domains. In: COLING 2270–2279

Morstatter F, Kumar S, Liu H, Maciejewski R (2013) Understanding twitter data with tweetxplorer. In: Proceedings of the 19th ACM SIGKDD International conference on Knowledge discovery and data mining 1482–1485

Neill DB, Gorr WL (2007) Detecting and preventing emerging epidemics of crime. Advances in Disease Surveillance 4:13

Neill DB, Wong WK (2009) Tutorial on Event Detection. KDD

Newswires (2015) https://www.newswire.com/, Accessed Feb 2016

Nurwidyantoro A, Winarko E (2013) Event detection in social media: a survey. In: ICT for Smart Society (ICISS). 2013 IEEE International Conference 1–5

Osborne M, Moran S, McCreadie R, Von Lunen A, Sykora M D, Cano E, Jackson T (2014) Real-time detection, tracking, and monitoring of automatically discovered events in social media. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. ACL 2014 37–42

Long R, Wang H, Chen Y, Jin O, Yu Y (2011) Towards effective event detection, tracking and summarization on microblog data. In: International Springer Conference on Web-Age Information Management 652–663

Papadopoulos et al. (2015) Social sensor. Report. http://www.socialsensor.eu/images/wp1_evaluation_report.pdf. Accessed Jan 2016

Papka R, Allan J (1998) On-line new event detection using single pass clustering title2. Technical Report. University of Massachusetts

Patwardhan S, Riloff E (2009) A unified model of phrasal and sentential evidence for information extraction. In: Proceedings of the 2009. Conference on Empirical Methods in Natural Language Processing of Association for Computational Linguistics 1:151–160

Pereira Nunes B, Mera A, Kawase R, Fetahu B, Casanova MA, de Campos GHB (2014) A topic extraction process for online forums. In: Advanced Learning Technologies (ICALT). 2014 IEEE 14th International Conference 541–543

Petrović S, Osborne M, Lavrenko V (2010) Streaming first story detection with application to twitter. In: Human Language Technologies. The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics 181–189

Popescu AM, Pennacchiotti M (2010) Detecting controversial events from twitter. In: Proceedings of the 19th ACM international conference on Information and knowledge management 1873–1876

Popescu AM, Pennacchiotti M, Paranjpe D (2011) Extracting events and event descriptions from twitter. In: Proceedings of the 20th ACM International conference companion on World Wide Web 105–106

Purohit H, Sheth AP (2013) Twitris v3: from citizen sensing to analysis, coordination and action. In: International Conference of Weblogs and Social Media (ICWSM) 2013 Jul

Qi Y, Candan KS (2006) Cuts: curvature-based development pattern analysis and segmentation for blogs and other text streams. In: Proceedings of the 17th ACM Conference on Hypertext and Hypermedia 1–10

Richardson M, Domingos P (2006) Markov logic networks. Mach Learn 62(1–2):107–136

Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th ACM International conference on World Wide Web 851–860

Sankaranarayanan J, Samet H, Teitler BE, Lieberman MD, Sperling J (2009) Twitterstand: news in tweets. In: Proceedings of the 17th ACM Sigspatial International conference on Advances in Geographic information systems 42–51

Sayyadi H, Hurst M, Maykov A (2009) Event detection and tracking in social streams. In: Proceedings of the 3rd International Conference of Weblogs and Social Media (ICWSM) 17–20

Schubotz T, Krestel R (2015) Online temporal summarization of news events. In: 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology 1:409–412

Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell 22(8):888–905

Song X, Tseng BL, Lin CY, Sun MT (2006) Personalized recommendation driven by information flow. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 509–516

Stewart A, Smith M, Nejdl W (2011) A transfer approach to detecting disease reporting events in blog social media. In: Proceedings of the 22nd ACM conference on Hypertext and Hypermedia 271–280

Stuart TL, Sandhu J, Stirling R, Corder J, Ellis A (2010) Campylobacteriosis outbreak associated with ingestion of mud during a mountain bike race. Epidemiol Infect 138(12):1695–1703

Tork H. (2011). Event Detection. Thesis. Laboratory of Artificial Intelligence and Decision Support (LIAAD-INESC TEC)

Trendsmap (2015) http://trendsmap.com/, Accessed January 2016

Tseng BL, Tatemura J, Wu Y (2005) Tomographic clustering to visualize blog communities as mountain views. In: WWW 2005 Workshop on the weblogging ecosystem

Twitter (2016) www.twitter.com, Accessed December 2015

Ushahidi (2008) https://www.ushahidi.com/, Accessed January 2016

Wan X, Milios E, Kalyaniwalla N, Janssen J (2009) Link-based event detection in email communication networks. In: Proceedings of the 2009 ACM symposium on Applied Computing 1506–1510

Wasi S, Shaikh ZA, Shamsi J (2011) Contextual event information extractor for emails. Sindh University Research Journal (SURJ) (Science Series), 43(1(a))

Weng J, Lee BS (2011) Event detection in twitter. Int Conf Weblogs Soc Media (ICWSM) 11:401–408

Wikipedia (2016) https://en.wikipedia.org/wiki/Wikipedia. Accessed 1November 2016

Xie Y (2011) Report on the public opinions and crisis management report. Social Science Literature Press, Beijing, 1–12 (in Chinese)

Xie W, Zhu F, Jiang J, Lim EP, Wang K (2013) Topicsketch: real-time bursty topic detection from twitter. In: 2013 IEEE 13th International Conference on Data Mining 837–846

Yang Y, Pierce T, Carbonell J (1998) A study of retrospective and on-line event detection. In: Proceedings of the 21st annual international ACM SIGIR Conference on Research and development in Information Retrieval 28–36

Yang Y, Carbonell JG, Brown RD, Pierce T, Archibald BT, Liu X (1999) Learning approaches for detecting and tracking news events. IEEE Intell Syst 14:32–43. doi:10.1109/5254.784083

Youtube (2016) www.youtube.com, Accessed December 2015

Zhao Q, Mitra P (2007). Event detection and visualization for social text streams. In: International Conference of Weblogs and Social Media. ICWSM

Zhao Q, Mitra P, Chen B (2007) Temporal and information flow based event detection from social text streams. In: Proceedings of the 22nd National Conference on Artificial Intelligence 2: 1501–1506

Zhao J, Wang X, Ma Z (2014) Towards events detection from microblog messages. Int J Hybrid Inf Technol 7(1):201–210

Zhi Li Wu, Chun Hung Li (2007) Topic detection in online discussion using non-negative matrix factorization. In: Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology Workshops 272–275

Zhou X, Chen L (2014) Event Detection over twitter social media streams. VLDB J 23(3):381–400. doi:10.1007/s00778-013-0320-3

Zhou D, Chen L, He Y (2014) A simple bayesian modelling approach to event extraction from twitter. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, USA 700–705

Zhu M, Hu W, Wu O (2008) Topic detection and tracking for threaded discussion communities. In: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology 1:77–83