Multimed Tools Appl
DOI 10.1007/s11042-015-2731-1

Crowdsourcing based social media data analysis of urban emergency events

Zheng Xu[1,2] · **Yunhuai Liu**[1] · **Junyu Xuan**[3] · **Haiyan Chen**[4] · **Lin Mei**[1]

Received: 24 August 2014 / Revised: 18 April 2015 / Accepted: 3 June 2015

Abstract An urban emergency event requires an immediate reaction or assistance for an emergency situation. With the popularity of the World Wide Web, the internet is becoming a major information provider and disseminator of emergency events and this is due to its real-time, open, and dynamic features. However, faced with the huge, disordered and continuous nature of web resources, it is impossible for people to efficiently recognize, collect and organize these events. In this paper, a crowdsourcing based burst computation algorithm of an urban emergency event is developed in order to convey information about the event clearly and to help particular social groups or governments to process events effectively. A definition of an urban emergency event is firstly introduced. This serves as the foundation for using web resources to compute the burst power of events on the web. Secondly, the different temporal features of web events are developed to provide the basic information for the proposed computation algorithm. Moreover, the burst power is presented to integrate the above temporal features of an event. Empirical experiments on real datasets show that the burst power can be used to analyze events.

Keywords Crowdsourcing · Social media analysis · Urban emergency events

1 Introduction

With the help of cloud computing [12, 20, 21, 31, 38], the internet of things [19, 27, 28], and Big Data [34, 35], crowdsourcing connects unobtrusive and ubiquitous sensing technologies, advanced data management and analytics models, and novel visualization methods, to create

✉ Zheng Xu
xuzheng@shu.edu.cn

[1] The Third Research Institute of the Ministry of Public Security, Shanghai, China

[2] Tsinghua University, Beijing, China

[3] Shanghai University, Shanghai, China

[4] East China University of Political Science and Law, Shanghai, China

Published online: 27 June 2015

solutions that improve the urban environment, human life quality, and city operation systems. An emergency event is a sudden, urgent, usually unexpected incident or occurrence that requires an immediate reaction or assistance for an emergency situation faced by a social group (e.g., a corporation) or recipients of emergency assistance [5, 9–11, 13, 22]. Therefore, how to prepare for, respond to, and recover from an emergency event is important. A method for processing an emergency event is to analyze its related information. Especially, with the development of the social media, people can get/post more and more information in terms of an emergency from/to the web. In fact, a web user can be seen as a sensor of an emergency event. For example, if a user makes a post in micro blogs or BBS about an earthquake occurrence, then she/he can be seen as an "earthquake sensor". The web can therefore be seen as a sensor receiver.

In this paper, the crowdsourcing based social media data analysis of urban emergency events is introduced. This is the foundation to detect urban emergency events using social media. Topic detection and tracking (TDT) traces the development of events. However, it is not a measure of the dynamic evolution process of the events, and does not consider event semantic characteristics in the evolution process. Indeed the TDT cannot offer us a global and clear understanding of the web event. In order to get beyond the insufficient nature of the TDT, we therefore propose a method to measure the evolutionary process of the urban emergency event based on burst and fluctuation power. Further, we put forward an approach to distinguish the event type based on fuzzy recognition. The major contributions of this paper are as follow.

(1) This paper proposes a crowdsourcing based burst computation algorithm of an urban emergency event in order to convey information about the event clearly and to help particular social groups or governments to process events effectively. The proposed method serves as the foundation for using web resources to compute the burst power of events on the web.
(2) The proposed model is based on crowdsourcing, which uses the real-time nature of Weibo users. The proposed model is applied into the emergency field, which can provide useful information to analyze and resist urban emergency events. The different type of emergency event is given and detected.
(3) Experiments on real data sets show the proposed method has good performance and high effectiveness in the analysis and describing of urban emergency events.

The rest of the paper is organized as follows. In the next section, the related work is given. Section 3 gives the problem formulation the proposed method. Section 4 presents a detailed method for computing the burst power. Section 5 gives the method for analyzing event type. Experiments on real emergency events are conducted in Section 5. The last section gives the conclusions of our work.

2 Related work

The proposed problem is similar to the research of Topic Detection and Tracking (TDT). Various methods have been proposed to manage news stories, spot news events, and track the process of events [1, 3, 6, 26, 32, 37]. Usually, the TDT research generates a hierarchical structure of an event, which aims at clustering the related news into it. Overall, TDT technologies have been attempting to detect or cluster news stories into these events, without

focusing on or interpreting the sudden, urgent, and unexpected features of emergency events [2].

Event evolution as proposed by Makkonen [23] is a subtopic of topic detection and tracking. In his study, two important conclusions are given: (1) a seminal event may lead to several other events; (2) the events at the beginning may have more influence on the events coming immediately after than the events at a later time. Makkonen used ontology to measure the similarity of events. However, ontology is difficult to get, which makes the work difficult to be used directly. Mei [30] investigated theme evolution which is similar to event evolution. He proposed a temporal pattern discovery technique on the basis of the timestamps of text streams. The theme of each interval is identified, and the evolution of the theme between successive intervals is extracted. Unfortunately, the proposed technique did not consider the different states of an event, which may impact on its result. Wei [36] proposed an event evolution pattern discovery technique which identifies event episodes together with their temporal relationships. An event episode is defined as a stage or sub-event of an event. The above study differs from this paper: their study deals with an event and event episodes, whereas this paper handles the different states of emergency events imaged on the web. Later, Yang [15] aimed at producing event evolution graphs from news corpora. The proposed event evolution graph is used to present the underlying structure of events. The proposed method used the event timestamp, event content similarity, temporal proximity, and web pages distribution proximity to model the event evolution relationships. Recently, Jo [24] studied a method to discover the evolution of topics (i.e., events) over time in a time-stamp document collection. He tried to capture the topology of topic evolution that is inherent in a given corpus. He claimed that the topology of the topic evolution discovered by his method is very rich and carries concrete information on how the corpus has evolved over time.

Event detection based on prior user queries is reported in [14, 25]. Fung et al. [25] proposed to first identify the bursty features related to the user query and then organize the documents related to those bursty features into an event hierarchy. In [14], a user specifies an event (or a topic) of interest using several keywords as a query. The response to the query is a combination of streams (e.g., news feeds, emails) that are sufficiently correlated and collectively contain all query keywords within a time period. The proposed work is also related to event detection using click-through data [39]. Event ranking with user attention is reported in [16] where the events are firstly detected from news streams. User attention is then derived from the number of page-views (collected through web browser toolbars) for all the news articles in the same event. Leskovec et al. [7, 21] proposed a method for outbreak detection based on cost-effective functioning.

Recently, with the high speed development of social networks such as Twitter and Weibo, many researchers have published their work of using data from social networks. Sakaki et al. [33] investigated the real-time nature of Twitter, and placed particular attention on event detection. The Twitter users are regarded as sensors. Their messages are used for detecting earthquakes. A reporting system is developed for use in Japan by their proposed methods. Crooks et al. [4] thought of Twitter as a distributed sensor system. They analyzed the spatial and temporal features of the Twitter feed activity responding to a 5.8 magnitude earthquake. Their experimental results argued that the Twitter users represent a hybrid form of a sensor system that allows for the identification and localization of the impact area of the event. Longueville et al. [8] used Twitter as a source of spatial-temporal information. By focusing on the real-life case of a forest fire, they aimed to demonstrate its possible role in support emergency planning, risk and damage assessment activities. Besides the emergency events

management, other researchers use the spatial and temporal information from social networks to support location based services. Liu et al. [40] presented MoboQ, the location-based real-time question answering service that is built on top of a microblogging platform. Qu and Zhang [29] used Twitter user generated mobile location data for trade area analysis. Their model includes three key processes: identifying the activity center of a user, profiling users based on their location history, and modeling users' preference probability.

3 Problems formulation

In this section, the basic definition of the proposed method is introduced. The temporal features of an urban emergency event are given in the next section. The burst factors of the proposed method are given in the last section.

3.1 Basic definitions

An event is something that happens at some specific time, and often some specific place. In fact, this definition of events from TDT can be relaxed since some events do not happen at some specific place. Many events are launched by some news or stories only on the web which cannot get the exact location stamp. However, the time of an emergency event can always be identified since some news of it has an exact timestamp t. Besides the timestamp, in this paper, we only take web pages into consideration since they can be easily processed and analyzed. An emergency event is defined as follows.

Definition 1 Emergency Event, *e*. An emergency event e is a tuple $\{L_e,F_e\}$, where L_e is the life course of e, F_e is the set of basic features describing e.

Definition 2 The basic features and life cycle of emergency event imaged on the web, F_e and L_e. The basic features of an emergency event contain three components including a seeds set $S(t_i,t_j)$, web pages set $\varphi(t_i,t_j)$, and keywords set $\psi(t_i,t_j)$. The life cycle of an emergency event contains the starting and ending timestamp $\langle t_s,t_e\rangle$.

A seeds set consists of the core keywords of an emergency event from timestamp t_i to t_j. Usually, these keywords can be used to search the related web pages covering one web event. A web page set is a set with n related web pages returned by search engines using the seeds from timestamp t_i to t_j, which is denoted as $\varphi(t_i,t_j)=\{d_1,d_2,\ldots,d_i,\ldots,d_n\}$. A keyword set is a set with m keywords extracted from $\varphi(t_i,t_j)$, which is denoted as $\psi(t_i,t_j)=\{k_1,k_2,\ldots,k_i,\ldots,k_m\}$. Web page d_i is represented by a keyword vector, which is denoted as

$$d_i = \{w_1, w_2, \ldots, w_m\} \tag{1}$$

where $w_j=(1+\log tf(k_j))*\log(1+n/df(k_j))$ [25]; $tf(k_j)$ means the term frequency of keyword k_j in web page d_i; $df(k_j)$ means the web page frequency of keyword k_j in $\varphi(t_i,t_j)$.

Herein, we use search engines such as Google[1] and Baidu[2] to get the web pages related to an emergency event imaged on the web. The reasons are as follows.

[1] www.google.com
[2] www.baidu.com

(1) Updating information rapidly. The information quickly refreshes an event. For example, the information of "Japan nuclear crisis" from social sensors may update per hour even per minute. Web search engines such as Google provide the interface to search information up to the minute.
(2) Different information sources. The information of an emergency event usually comes from different sources such as news, blogs, and bbs. Web search engines such as Google and Baidu provide the interface to search information from various websites.

3.2 Basic temporal features

In this section, we presents five basic temporal features including: 1) the number of increased web pages, 2) the number of increased keywords, 3) the distribution of keywords on web pages, 4) the associated relations of keywords, and 5) the similarities of web pages.

Temporal Feature 1 The number of increased web pages from timestamp t_i to t_j, $|\varphi(t_i,t_j)|$. The elements in $\varphi(t_i,t_j)$ do not appear from the starting timestamp t_s to t_i, that is, $\forall d_n \in \varphi(t_i,t_j) \rightarrow d_n \notin \varphi(t_s,t_i)$.

Temporal Feature 2 The number of increased keywords from timestamp t_i to t_j, $|\psi(t_i,t_j)|$. The elements in $\psi(t_i,t_j)$ do not appear from the starting timestamp t_s to t_i, that is, $\forall k_m \in \psi(t_i,t_j) \rightarrow k_m \notin \psi(t_i,t_j)$.

Temporal Feature 3 The distribution of keywords on web pages from timestamp t_i to t_j, $\zeta(t_i, t_j)$. For an emergency event e, the web pages in $\varphi(t_i,t_j)$ can be represented as a vector by the keywords in $\psi(t_i,t_j)$. These vectors can be stored as a matrix:

$$\zeta(t_i,t_j) = \begin{pmatrix} w_{11} & \cdots & w_{1m} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nm} \end{pmatrix}. \tag{2}$$

Temporal Feature 4 The associated relations of keywords from timestamp t_i to t_j, $\Gamma(t_i,t_j)$. For an emergency event e, the associated relations of keywords can be stored as a matrix:

$$\Gamma(t_i,t_j) = \begin{pmatrix} f_{11} & \cdots & f_{1m} \\ \vdots & \ddots & \vdots \\ f_{m1} & \cdots & f_{mm} \end{pmatrix}. \tag{3}$$

where f_{ij} means the weight of relation between k_i and k_j, which can be computed by

$$f_{ij} = \frac{\log\left(\frac{N(k_i \wedge k_j) * n}{N(k_i) * N(k_j)}\right)}{\log n} \tag{4}$$

where $N(k_i)$ means the number of in $\varphi(t_i,t_j)$ containing k_i; $N(k_i \wedge k_j)$ is the number of web pages in $\varphi(t_i,t_j)$ containing both k_i and k_j.

Temporal Feature 5 The similarities between web pages from timestamp t_i to t_j, $\Xi(t_i,t_j)$. For an emergency event e, the similarities between web pages can be stored as a matrix:

$$\Xi(t_i,t_j)=\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}. \tag{5}$$

where a_{ij} means the similarities between d_i and d_j, which can be computed by

$$a_{ij}=\frac{d_i \cdot^i d_j}{\|d_i\|\|d_j\|}. \tag{6}$$

where $\|d_i\|$ and $\|d_j\|$ denote the mathematical model of vector d_i and d_j.

3.3 Basic burst factors

In this section, we present basic burst factors including: the number of communities in a context graph, the average clustering coefficient of the context graph, and the average similarities of web pages.

Impact Factor 1 The number of communities in a context graph from timestamp t_i to t_j, $|C(t_i,t_j)|$. A community is a subgraph of the context graph, which reflects a part of the context of an event e. The set of communities is a segmentation of the context graph. Each context community is a part of the context graph, which has no common keywords of other communities. The set of communities of an event e is denoted as:

$$\begin{aligned} &C_e=\{c_1,c_2,\ldots,c_{|C_e|}\} \\ &\forall k_i \in c_i \wedge k_j \in c_j \rightarrow k_i \neq k_j \end{aligned} \tag{7}$$

Impact Factor 2 The average clustering coefficient [33] of the context graph from timestamp t_i to t_j, $CC(t_i,t_j)$. In graph theory, a clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. The clustering coefficient of the keyword k_i in a context graph can be computed by

$$CC(k_i)=\frac{2l}{p(p-1)} \tag{8}$$

where p means the number of neighbor nodes of the keyword k_i and l means the number of edges between these neighbor nodes. Thus, the average clustering coefficient of the context graph can be computed by

$$CC(t_i,t_j)=\frac{\sum_{i=1}^{m} CC(k_i)}{m} \tag{9}$$

Impact Factor 3 The average similarities of web pages from timestamp t_i to t_j, $AS(t_i,t_j)$. For an event e, the similarities can be computed by the cosine function

$$Sim(d_i, d_j) = \frac{d_i \cdot^i d_j}{\|d_i\| \|d_j\|} \tag{10}$$

where $\|d_i\|$ and $\|d_j\|$ denote the mathematical model of vector d_i and d_j. Thus, the average similarities of web pages can be computed by

$$AS(t_i, t_j) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} Sim(d_i, d_j)}{m(m-1)} \tag{11}$$

4 Computing the burst power

In this section, based on the above definitions, the method for computing the burst power of an emergency event is proposed.

4.1 Basic definitions of burst power

After giving the basic temporal features of emergency events, we define burst power as follows.

Definition 7 Burst Power, $op(t_i,t_j)$. For an emergency event e, the burst power from timestamp t_i to t_j is the influence degree to society.

For example, high $|\varphi(t_i,t_j)|$ or $|\psi(t_i,t_j)|$ means a high influence degree of an event to society, thus the event has high burst power.

According to the characteristics of burst power, the time interval with a high influence to society will possess a high possibility to be the peak or milestone of an emergency event. Inspired by [39], before we propose the algorithm to compute the burst power, we will introduce two important definitions as follows.

Definition 8 The representative power of keywords, $rp(k)$. The representative power of keyword k is the probability of k to represent the event e correctly.

Definition 9 The confidence of web pages, $cw(d)$. The confidence of a web page d is the expected representative power of keywords provided by d.

Different keywords related with one event reveal the various aspects of the event. For example, the keyword "China" reveals the place of the event "China rail crash". On the other hand, the keyword "rail" and "crash" reveals the object of the event "China rail crash".

4.2 Basic heuristics for computing burst power

Based on common sense and the observations on real data, we have four basic heuristic rules which serve as the bases for the computing of burst power. These four heuristic rules are relevant to the data. If there is a burst situation in the data, they will be correct. Given the

discussion field of this paper, all emergency events have different effects and spreading powers. Thus, these four heuristic rules are appropriate for emergency events situations.

Heuristic rule 1 If ignoring $\psi(t_i,t_j)$ and $\zeta(t_i,t_j)$, the possibility of time interval (t_i,t_j) with high burst power increases with $|\varphi(t_i,t_j)|$.

According to heuristic rule 1, the burst power is proportional to the number of increased web pages. So we can get, $op(t_i,t_j)\infty|\varphi(t_i,t_j)|$.

Heuristic rule 2 If ignoring $\varphi(t_i,t_j)$ and $\zeta(t_i,t_j)$, the possibility of a time interval (t_i,t_j) with high burst power increases with $|\psi(t_i,t_j)|$.

According to heuristic rule 2, the burst power is proportional to the number of increased keywords. So we can get, $op(t_i,t_j)\infty|\psi(t_i,t_j)|$.

If two time intervals with the same $|\varphi(t_i,t_j)|$ and $|\psi(t_i,t_j)|$, the distribution of keywords will determine $op(t_i,t_j)$. So, we give heuristic rule 3.

Heuristic rule 3 If the bipartite graph of $\zeta(t_i,t_j)$ is a complete graph, $op(t_i,t_j)$ is the lowest, that is, $(\forall w_{nm}\in\zeta(t_i,t_j)\rightarrow w_{nm}\neq 0)\rightarrow op(t_i,t_j)_{\min}$.

According to heuristic rule 3, if all of the keywords appear in each web page, the similarity between them is 1, which means all of the web pages are copies from one web page. This situation means that the emergency event is that with the lowest diversity.

Since $\psi(t_i,t_j)$ and $\varphi(t_i,t_j)$ are dependent, the distribution of keywords on the web pages should be considered. So, we give heuristic rule 4.

Heuristic rule 4 Since $\psi(t_i,t_j)$ and $\varphi(t_i,t_j)$ are dependent, the distribution of keywords on web pages should be considered. If all of the keywords are provided by only one web page, $op(t_i,t_j)$ is the highest.

4.3 Computing the burst power

Heuristic rule 4 gives the condition of maximum $op(t_i,t_j)$. Figure 1 gives the illustration of heuristic rule 3 and 4. Based on heuristic rule 3 and 4, we can conclude that the confidence of a

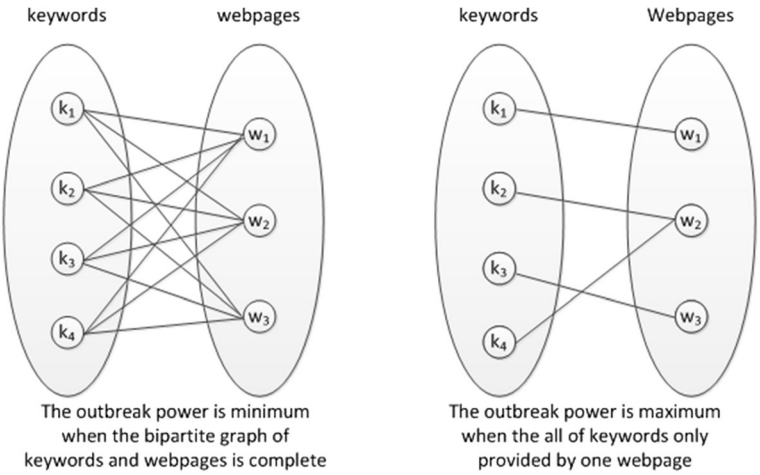


Fig. 1 The illustration of the maximum and minimum burst power

web page and the representative power of a keyword are determined by each other, and we can use an iterative method to compute them. Thus, we compute the confidence of a web page by calculating the average representative power of keywords that it provides as follows:

$$cw(d_h) = \frac{\sum_{\left(k_m \in \vec{d_h}\right) \bigcap (w_{hn} > 0)} rp(k_m)}{\left|\vec{d_h}\right|} \tag{12}$$

where $\left|\vec{d_h}\right|$ means the number of web pages with keywords k_m.

Inspired by [39], we use a probability function to compute the representative power of keywords:

$$rp(k_m) = 1 - \prod_{\left(d_h \in \vec{k_m}\right) \bigcap (w_{hm} > 0)} (1 - cw(d_h)) \tag{13}$$

where $\vec{k_m}$ is the set of web pages providing k_m.

The above two equations show how to compute the confidence of a web page. However, since the similarities between web pages are not zero, we put the similarities between web pages into Eq. 8. The equation can be revised as Eq. 9, which considers the similarities between web pages:

$$rp'(k_m) = rp(k_m) + \sum_{\left(k_i \in \vec{kk_i}\right) \bigcap (f_{ij} > 0)} f_{ij} * rp(k_i) \tag{14}$$

where $\vec{kk_i}$ is the set of keywords similarities against keyword k_i.

Since $rp'(k)$ may be higher than 1, we adopt the widely used logistic function to set $rp'(k)$ into (0, 1). Then Eq. 9 can be revised as

$$rp''(k_m) = \frac{1}{1 + e^{-rp'(k_m)}} \tag{15}$$

Equation 7 is revised as:

$$cw(d_h) = \frac{\sum_{k_m \in \vec{d_h} \bigcap w_{hm} > 0} rp''(k_m)}{\left|\vec{d_h}\right|} \tag{16}$$

and the burst power function of time interval (t_i, t_j) can be computed by the sum of the confidence of each web page:

$$op(t_i, t_j) = \sum_{h=1}^{n} (1 - cw(d_h)) \tag{17}$$

where n means the number of web pages from time interval (t_i, t_j).

As described above, we can infer the confidence of a web page if we know the representative power of keywords and vice versa. The algorithm for computing the burst power is given by algorithm 1.

The steps in algorithm 1 try to compute the burst power and representative power, and it stops when the computation reaches a stable state. As in other iterative approaches, we choose the initial state in which all web pages have uniform confidence α.

5 Distinguish the event type based on the fuzzy recognition

5.1 The web event type

Definition 10 Web event type: ε.
$E=\{e_1,e_2,\ldots,e_s\}$. represents the web event set. $\varepsilon=\{\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n\}$, $\varepsilon_i \subseteq E$ represents the web event type set. ε_i is the one of the web event types, and n is the number of the web event types. According to the observation of the actual data set and the cognitive sense, and comprehensive considering variety factors (the nature of the web event, severity, controllability and the influence range), we think web events can be classified into three kinds, namely: $\varepsilon=\{\varepsilon_1,\varepsilon_2,\varepsilon_3\}$. ε_1 represents the Urgent events, ε_2 represents the Hot events ε_3 represents the General events.

First, we need to extract some important features from the attained time-series data of the semantic outbreak power. Make the time-series data of the semantic outbreak power $S=\{s_1,s_2,\ldots,s_n\}$ in its full life cycle L_e. s_i represents the semantic outbreak power of the time *i.* n is the length of time series. We calculate the related parameters in order to study its characteristics of the web events.

1) Average outbreak power: op_{ave}

$$op_{ave} = \frac{1}{n}\sum_{i=1}^{n} s_i \tag{18}$$

Average outbreak power op_{ave} reflects the average heat or attention of the events in its full life cycle L_e.

2) Fluctuation Power: *fp*

$$fp = \frac{Var}{op_{ave}}, \quad Var = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(s_i - op_{ave})^2} \tag{19}$$

Fluctuation Power *fp* reflects the volatility of the events semantic outbreak power outbreak power in its full life cycle L_e. Then, we define the events type comprehensively considering their own characteristics.

Definition 11 Urgent events: ε_1.

The web urgent events refer to the major natural disasters, accidents disasters or social safety incidents which need to be handled by the state or the government. For example: "5.12 wenchuan earthquake", "7.23 the wenzhou crash", "9.11 terrorist attacks" and so on. These events generally have the following features (a) sudden and complex; (b) destructive and sustainable; (c) the transmission range is larger; (d) they cause extensive societal concern.

According to the observation of the semantic outbreak power of urgent events, we find that the web urgent events usually have a relatively modest average outbreak power op_{ave} and are accompanied by a relatively large Fluctuation Power, namely:

$$\varepsilon_1 = \left\{ \left\{ E' \middle| E' \subseteq E \wedge \left(\forall e \in E' \right\} \rightarrow (\delta_2 < op_{ave}(e) \leq \delta_1 \ \&\& \ fp(e) > \theta_1) \right\}$$

Definition 12 Hot events: ε_2.

The hot issues on the web refer to the daily focused topics on the web or in the real world, such as: "price control", "food safety", etc. The media will report on and discuss these events. However, the possibility of these events causing harm to individuals, groups or society is lower. These events have the following characteristics: (a) the uncertainty of event attributes; (b) different web page about these events have a certain similarity; (c) these events have the possibility to evolve. According to the observation of the semantic outbreak power of hot events, we find that the web hot events usually have a relatively high average outbreak power op_{ave} and are accompanied by a relatively modest Fluctuation Power, namely:

$$\varepsilon_2 = \left\{ E' \middle| E' \subseteq E \wedge \left(\forall e \in E' \right) \rightarrow (op_{ave}(e) > \delta_1 \ \&\& \ \theta_2 < fp(e) \leq \theta_1) \right\}$$

Definition 13 General events: ε_3.

General events on the web refer to events which generate less attention on the web or in the real world. These events don't cause harm to individuals, groups or society. These events have the following characteristics: (a) the certainty of event attributes; (b) the similarity of different web pages about these events is 1; (c) the evolution of these events is slower. According to the observation of the semantic outbreak power of hot events, we find that the general web events usually have a relatively low average outbreak power op_{ave} and are accompanied by a relatively low Fluctuation Power, namely:

$$\varepsilon_3 = \left\{ E' \middle| E' \subseteq E \wedge \left(\forall e \in E' \right) \rightarrow (op_{ave}(e) \leq \delta_2 \ \&\& \ fp(e) \leq \theta_2) \right\}$$

5.2 Distinguish the event type based on the fuzzy recognition

Here, we give an algorithm to distinguish the event type based on fuzzy recognition. We can use a set of feature vectors to represent an event. For a web event e, one set of feature vectors can be attained from the time-series data of the semantic outbreak power, namely: $\vec{T_d}(e) = \{x_1, x_2, \ldots, x_m\}$;where x_j is the j and the feature component for the event e. Obviously, every feature x_j can be directly obtained from the time-series data of semantic outbreak power, and different event types ε_i have different patterns in terms of features. Therefore, we need to establish the contact and mapping from the event features to the event types. Inspired by the [i] [ii], this paper applies fuzzy mathematics theory to identify types of web events. We use the membership $\xi_{ij}(x_j)$ to measure the power of the event feature x_j belonging to a certain power of an events ε_i. After statistics determines the known web event data set (the training data set with artificial tagging event types), we can calculate the membership distribution of the event ε_i when the feature x_j has a

different value. Then, we got the events membership function vector by gathering every membership distribution of each feature together, namely:

$$\vec{\varepsilon_i} = \{\xi_{i1}, \xi_{i2}, \ldots, \xi_{im}\} \tag{20}$$

After establishing the prior knowledge of distinguishing the event type, we can use prior knowledge to judge unknown web event types. For a web event e, the features vector is $\vec{T_d}(e) = \{x_1, x_2, \ldots, x_m\}$, then the membership vector of the event e belongs to the ith type event ε_i which is:

$$\vec{\varepsilon_i}(e) = \{\xi_{i1}(x_1), \xi_{i2}(x_2), \ldots, \xi_{im}(x_m)\} \tag{21}$$

Where $\xi_{ij}(x_j)$ is the membership of the jth feature component belonging to the ith type for the event e.

Then, the membership of the event belonging to the ith type is:

$$\varepsilon_i(e) = \sum_{j=1}^{m} \xi_{ij}(x_j) * \alpha_j, \sum_{j=1}^{m} \alpha_j = 1 \tag{22}$$

Where α_j is the weight factor of the jth feature component? Assuming that the different features components are of the same importance, namely:

$$\varepsilon_i(e) = \frac{1}{m} \sum_{j=1}^{m} \xi_{ij}(x_j) \tag{23}$$

Next, we give the following deduction as the algorithm reason to judge the event type based on the maximum membership principle in the theory of fuzzy mathematics [17, 18].

6 Experiments and analysis

6.1 Data sets

The events in our experiments are extracted from Google and Baidu. We select 150 events with about 1,500,000 web pages in our experiments including political events, accidents events, disasters events, terrorism events, etc. The web pages of each event are downloaded from Google. Stanford tagger[3] is used to reserve the noun words in the web pages. The keywords are selected by their document frequencies. Table 1 shows the statistics of our experimental data set. When Google and Baidu provide the events, they also give some keywords for helping users to search them. After we get the seed set of an event by the search engine, a certain number of the web pages are collected as samples by automatic crawling and searching with the seed set. The detailed steps for collecting related web resources of an event in our experiments are as follows.

(1) Get the seed set of an event such as "China train crash", which can be seen as the $S(t_i, t_j)$ of an event;

[3] Nlp.stanford.edu.com

Table 1 The details of data sets

Feature	Value
Average number of seeds per event	2
Average number of web pages per event	1012
Average number of keywords per event	4534
Average number of days per event	30
Average number of web pages per day	34
Average number of keywords per day	853

(2) Search the seed set as the query, download the related web pages with a search engine, which can be seen as the $\varphi(t_i,t_j)$ of an event;
(3) Identify the starting timestamp of an event by $\varphi(t_i,t_j)$, and the ending timestamp by its download time, which can be seen as the t_s and t_e of an event;
(4) Get $|\varphi(t_i,t_j)|$, $|\psi(t_i,t_j)|$, $\zeta(t_i,t_j)$, and $\Xi(t_i,t_j)$ per day.
(5) Do step (4) of different information sources including news, blogs, and bbs.

6.2 Experimental results

After obtaining the temporal features per day of each event, we select ten human annotators to test whether the burst power of each event is correct or not. The data from Google Trends is selected to compare with the proposed method. If the human annotator thinks the proposed methods are equal to his own imagination or Google Trends, we set the precision of this detection to true. In addition, the annotators are set to do their evaluations independently, which ensures the reliability and validity of the results. Before the human annotator starts to evaluate the experimental results, we provide them with the abstract descriptions of each event. For example, we give them some news stories and concepts. This training session continues until the human annotator is familiar with the concepts and the temporal feature of events. In the next step, the annotator is asked to test the results of each event independently.

In fact, the overall precision of the proposed method achieves 92 %, showing the accuracy of our burst power computation algorithm. From the experiments on the real data we know that the proposed algorithm can compute the burst power of an event accurately. The information from the web can be integrated to the burst power. The factor can thereby be used to detect the various states of an emergency event.

6.3 The correctness verification of the event type distinguishes the web event

We choose the 100 web event data set as the experimental subject, of which 60 events constitute the training set and the rest represent the test set. The detailed results are shown in Table 2. The 60 events of the training set manually tag each event type. We choose appropriate feature value intervals and the statistics of the training set, then we can calculate the membership of every web event type when the event feature has different values. Figure 2 shows the details of the membership distribution when the outbreak power has different values. Figure 3 shows the details of the membership distribution when the fluctuation power has different values.

Table 2 Accuracy and recall of all kinds of Web event discriminant

Evaluation indicator	The Web event type			The total of all Web events
	Urgent	Hot	General	
Recall	95.2 %	88.8 %	90 %	91 %
Precision	90.9 %	88.8 %	85 %	88 %

We obtain the membership functions of all kinds of events by appropriate function fitting of the membership distribution. The formula (10) and (11) respectively show the membership of the features of the outbreak of fluctuation power about all of the different events.

$$\begin{aligned}
\xi_3(op_{ave}) &= \begin{cases} 1, & op_{ave} \le 0.575 \\ \dfrac{0.675 - op_{ave}}{0.675 - 0.575}, & 0.575 < op_{ave} \le 0.675 \\ 0, & op_{ave} > 0.675 \end{cases}, \\
\xi_2(op_{ave}) &= \begin{cases} 0, & op_{ave} \le 0.575 \\ 20op_{ave}^2 - 24op_{ave} + 7.2, & 0.575 < op_{ave} \le 0.825 \\ 0.6, & op_{ave} > 0.825 \end{cases}, \\
\xi_1(op_{ave}) &= \begin{cases} 0, & op_{ave} \le 0.575 \\ -52op_{ave}^2 + 73op_{ave} - 25, & 0.575 < op_{ave} \le 0.825 \\ 0.4, & op_{ave} > 0.825 \end{cases}
\end{aligned} \tag{24}$$

$$\begin{aligned}
\xi_3(op_{ave}) &= \begin{cases} 1, & op_{ave} \le 0.575 \\ \dfrac{0.675 - op_{ave}}{0.675 - 0.575}, & 0.575 < op_{ave} \le 0.675 \\ 0, & op_{ave} > 0.675 \end{cases}, \\
\xi_2(op_{ave}) &= \begin{cases} 0, & op_{ave} \le 0.575 \\ 20op_{ave}^2 - 24op_{ave} + 7.2, & 0.575 < op_{ave} \le 0.825 \\ 0.6, & op_{ave} > 0.825 \end{cases}, \\
\xi_1(op_{ave}) &= \begin{cases} 0, & op_{ave} \le 0.575 \\ -52op_{ave}^2 + 73op_{ave} - 25, & 0.575 < op_{ave} \le 0.825 \\ 0.4, & op_{ave} > 0.825 \end{cases}
\end{aligned} \tag{25}$$

Following this, we use the test set to verify this prior knowledge. Table 2 shows the statistical results. The precision of the total of all web events is 88 %, and the recall rate is 91 %. The test results show that the algorithm has good performance.

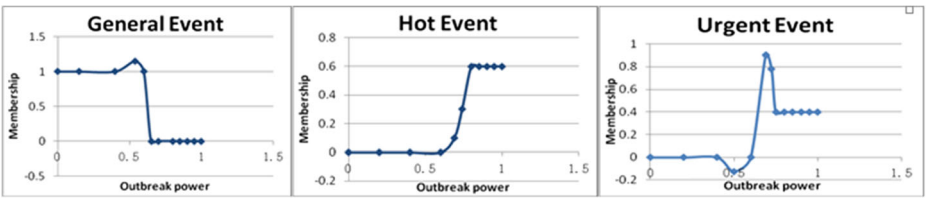


Fig. 2 The event "Libya unrest" evolution based on semantic outbreak power

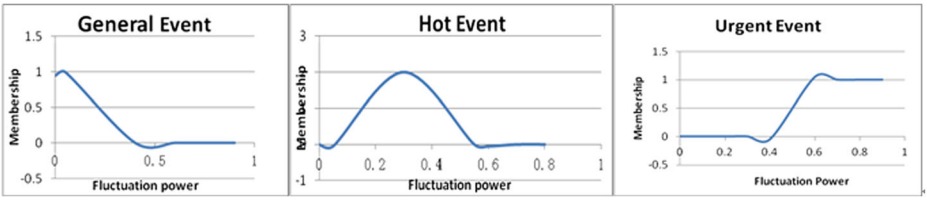


Fig. 3 The event "Libya unrest" evolution based on semantic fluctuation power

7 Conclusions

With the popularity of web, the internet is becoming a major information provider and a disseminator of events due to its real-time, open, and dynamic features. However, faced with huge, disordered and continuous web resources, it is impossible for people to efficiently recognize, collect and organize events. In this paper, a crowd sensing based burst computation algorithm of a web event is developed in order to clearly inform people about a web event and to help the particular social group or government process the event effectively. The definition of urban emergency event is introduced. This is the foundation of using web resources to compute the burst power of events on the web. Secondly, different temporal features of web events are developed to provide the basic features of the proposed computation algorithm. Moreover, the burst power is presented as a means to integrate the above temporal features of an event. Empirical experiments on real datasets show that the number of web pages and the average clustering coefficient can be used to detect events.

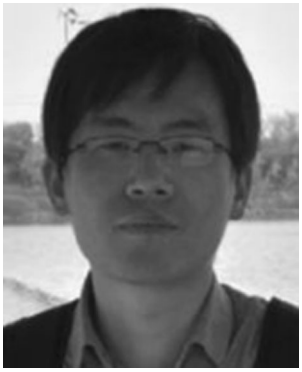
Acknowledgments This work was supported in part by the National Science and Technology Major Project under Grant 2013ZX01033002-003, in part by the National High Technology Research and Development Program of China (863 Program) under Grant 2013AA014601, 2013AA014603, in part by National Key Technology Support Program under Grant 2012BAH07B01, in part by the National Science Foundation of China under Grant 61300202, 61300028, in part by the Project of the Ministry of Public Security under Grant 2014JSYJB009, in part by the China Postdoctoral Science Foundation under Grant 2014 M560085, the project of Shanghai Municipal Commission of Economy and Information under Grant 12GA-19, and in part by the Science Foundation of Shanghai under Grant 13ZR1452900.

References

1. Abonyi J, Feil B, Nemeth S, Arva P (2005) Modified Gath–Geva clustering for fuzzing segmentation of multivariate time-series. Fuzzy Sets Syst Data Min 149:39–56
2. Allan J (2000) Topic detection and tracking: event-based information organization. Kluwer, Norwell
3. Allan J, Carbonell G, Doddington G, Yamron J, Yang Y (1998) Topic Detection and Tracking Pilot Study Final Report. In Proceedings of the Broadcast News Transcription and Understanding Workshop
4. Fung C, Yu X, Liu H, Yu S (2007) Time-dependent event hierarchy construction. In Proc. of KDD, pp 300–309
5. Haddow D, Bullock A, Coppola P (2010) Introduction to Emergency Management
6. He Q, Chang K, Lim E, Banerjee A (2010) Keep it simple with time: a reexamination of probabilistic topic detection models. IEEE Trans Pattern Anal Mach Intell 32(10):1795–1808
7. Himberg J, Korpiaho K, Mannila H, Tikanmaki J, Toivonen T (2001) Time series segmentation for context recognition in mobile devices. In Proceedings of the 2001 I.E. International Conference on Data Mining
8. Hristidis V, Valdivia O, Vlachos M, Yu PS (2006) Continuous keyword search on multiple text streams. In Proc of CIKM, pp 802–803
9. http://definitions.uslegal.com/e/emergency-event/, 2012

10. http://en.wikipedia.org/wiki/September_11_attacks, 2012
11. http://www.who.int/csr/sars/en/, 2012
12. Hu C, Xu Z et al Semantic Link Network based Model for Organizing Multimedia Big Data. IEEE Trans Emerg Top Comput doi:10.1109/TETC.2014.2316525
13. Jess F, Trina M, Jarrod T et al (2012) Riskr: A low-technological Web2.0 disaster service to monitor and share information. In Proceedings of 15th International Conference on Network-Based Information Systems, pp 1–8
14. Jin X, Spangler S, Ma R, Han J (2010) Topic Initiator Detection on the World Wide Web. In Proceedings of the 19th international World Wide Web conference, pp 481–490
15. Jo Y, Lagoze C, Lee Giles C (2007) Detecting research topics via the correlation between graphs and texts. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 370–379
16. Keogh E, Chakrabarti K, Pazzini M, Mehrotra S (2000) Dimensionality reduction for fast similarity search in large time series databases. J Knowl Inf Syst 3(3)
17. Leskovec J, Horvitz E (2008) Planetary-Scale views on a large instant-messaging network. In Proceedings of the 17th international World Wide Web conference
18. Leskovec J, Krause A, Guestrin C, Faloutsos C, VanBriesen J, Glance N (2007) Cost-effective outbreak detection in networks. In Proc. of KDD
19. Liu X, Yang Y, Yuan D, Chen J (2013) Do we need to handle every temporal violation in scientific workflow systems. ACM Trans Softw Eng Methodol
20. Liu Y, Zhu Y, Ni LM, Xue G (2011) A reliability-oriented transmission service in wireless sensor networks. IEEE Trans Parallel Distrib Syst 22(12):2100–2107
21. Luo X, Xu Z, Yu J, Chen X (2011) Building association link network for semantic link on web resources. IEEE Trans Autom Sci Eng 8(3):482–494
22. Makkonen J (2003) Investigation on event evolution in TDT. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language, pp 43–48
23. Mei Q, Zhai C (2005) Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp 198–207
24. Nallapati R, Feng A, Peng F, Allan J (2004) Event threading within news topics. In Proceedings of the thirteenth ACM international conference on Information and knowledge management, pps 446–453
25. Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Inf Process Manag 24(5):513–523
26. Sung C, Kim Collaborative T (2012) Modeling process for development of domain-specific discrete event simulation systems. IEEE Trans Syst Man Cybern Part C Appl Rev 42(4):532–546
27. Tang J, Wang M, Hua X-S, Chua T-S (2012) Social media mining and search. Multimed Tools Appl 56(1):1–7
28. Wang L, Tao J et al (2013) G-Hadoop: MapReduce across distributed data centers for data-intensive computing. Futur Gener Comput Syst 29(3):739–750
29. Wang C, Zhang M, Ru L, Ma S (2008) Automatic online news topic ranking using media focus and user attention based on aging theory. In Proc of CIKM, pp 1033–1042
30. Wei C, Chang Y (2007) Discovering event evolution patterns from document sequences. IEEE Trans Syst Man Cybern Part A 37(2):273–283
31. Wei X, Luo X, Li Q, Zhang J, Xu Z (2015) Online comment-based hotel quality automatic assessment using improved fuzzy comprehensive evaluation and fuzzy cognitive map. IEEE Trans Fuzzy Syst 23(1):72–84
32. Wu X, Lu Y, Peng Q, Ngo C (2011) Mining event structures from web videos. IEEE Multimedia 18(1):38–51
33. Xiong P, Fan Y, Zhou M (2009) Web service configuration under multiple quality-of-service attributes. IEEE Trans Autom Sci Eng 6(2):311–4321
34. Xu Z, Luo X, Zhang S, Wei X, Mei L, Hu C Mining Temporal Explicit and Implicit Semantic Relations between Entities using Web Search Engines. Fut Generat Comput Syst. doi:10.1016/J.future.2013.9.027
35. Xu Z et al Knowle: a Semantic Link Network based System for Organizing Large Scale Online News Events. Fut Generat Comput Syst. 10.1016/j.future.2014.04.002
36. Yang C, Shi X (2006) Discovering event evolution graphs from newswires. In Proceedings of the 15th international World Wide Web conference, pp 945–946
37. Yang C, Shi X, Wei C (2009) Discovering event evolution graphs from news corpora. IEEE Trans Syst Man Cybern—Part A 39(4):850–863
38. Yen NY, Zhang C, Waluyo AB, Park JJ (2015) Social media services and technologies towards web 3.0. Multimed Tools Appl. doi:10.1007/s11042-015-2461-4
39. Yin X, Han J, Yu PS (2008) Truth discovery with multiple conflicting information providers on the web. IEEE Trans Knowl Data Eng 20(6):796–808

40. Zhao Q, Liu T-Y, Bhowmick SS, Ma W-Y (2006) Event detection from evolution of click-through data. In Proc of KDD, pp 484–493



Zheng Xu was born in Shanghai, China. He received the Diploma and Ph.D. degrees from the School of Computing Engineering and Science, Shanghai University, Shanghai, in 2007 and 2012, respectively. He is currently working in the third research institute of ministry of public security and the postdoctoral in Tsinghua University, China. His current research interests include crowdsourcing, semantic Web and Web mining. He has authored or co-authored more than 50 publications.

Yunhuai Liu is a professor in the third research institute of ministry of public security, China. He received the PhD degrees from Hong Kong University of Science and Technology (HKUST) in 2008. His main research interests include wireless sensor networks, pervasive computing, and wireless network. He has authored or co-authored more than 50 publications and his publications have appeared in IEEE Trans. on Parallel and Distributed Systems, IEEE Journal of Selected Areas in Communications, IEEE Trans. on Mobile Computing, IEEE Trans. on Vehicular Technology etc.

Junyu Xuan is now a Ph. D student at University of Technology, Sydney. His current research interests include semantic Web and Web mining.

Haiyan Chen is an assistant professor in East China University of Political Science and Law, Shanghai, China. His current research interests include cloud computing.