Journal of Web Engineering, Vol. 15, No.3&4 (2016) 226-248

OUTBREAK POWER MEASUREMENT FOR EVOLUTION COURSE OF WEB EVENTS

XINZHI WANG

Shanghai University, Shanghai, China

Tsinghua University, Beijing, China

wxz15@mails.tsinghua.edu.cn

XIANGFENG LUO

Shanghai University, Shanghai, China

luoxf@shu.edu.cn

HUI ZHANG

Tsinghua University, Beijing, China

zhhui@mail.tsinghua.edu.cn

ZHENG XU

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

Tsinghua University, Beijing, China

xuzheng@shu.edu.cn

HUIMIN LIU

Shanghai University, Shanghai, China

hliu@shu.edu.cn

Received October 9, 2015
Revised December 31, 2015

Nowadays, emergencies have a great impact on people's daily lives. Web makes it possible to study emergencies from web information due to its real-time, open, and dynamic features. Measuring temporal features in web events evolution course can help people timely get knowledge and understand emergent events, which contribute to reducing harms to our society caused by emergencies. In this paper, we propose an outbreak power measuring algorithm for the evolution of web events, in order to provide guidance for automatic detection and prediction of emergencies. An iterative algorithm is firstly introduced to calculate outbreak power of web events through increased web pages of events, increased attributes of events, and distribution of attributes in web pages and the relationships of attributes. Secondly, definition of web events types is proposed. From studying each type of web events, we dig out feature patterns and find laws of each type events, with hot event having the highest outbreak power while general event have the lowest outbreak power, and general event fluctuating most while urgent event fluctuating least, which can be prior knowledge of web events we study. And then, a fuzzy based algorithm is presented to discriminate the type of web events. By means of prior knowledge, membership grade of web events

belong to each type can be calculated, and then the type of web events can be discriminated. Experiments on real data set demonstrate the proposed algorithm is both efficient and effective, and it is capable of providing accurate results of discrimination.

Key words: emergent events, web events, semantic measure, web mining, fuzzy pattern recognition
Communicated by: M. Gaedke & Q. Li

1 Introduction

Web event is what social Medias (i.e., BBS, blog, and news sites) discuss via cyber and influence on our real society. People can discuss web event in various forms, such as commenting news, posting and replying in forum, or recording and messaging in blog, etc. These discussions, which describe lots web events, have an impact on the evolution of web event. In return, our society will be influenced by the information of web. So the detection and prediction of web events evolution is a meaningful work. To get this goal, we make our efforts to measure and analyse evolution features of web events.

In today's world, emergencies [1, 2] have a great impact on people's daily lives. For example[3], on July 4, 2011, <News of the World> was revealed wiretapping the phone between missing girl Millie Dowler and her families in 2002 which stock the work of police. The tapping phone event caused a great feedback in British, and then a succession of eavesdropping scandals was reported. The results shocked the world. That scandal spread and reported by Web and media, which strained the relationship between media and politicians, which made citizens worry about their privacy. In other words, the tapping phone outbreak with lots sub event at last. Another example [4] happened in India in July 2012 is caused by a series of rumours and threats spread by SMS, web and other society media, leading to panic in the entire region. Text in the messages implied that local Muslim would begin a large-scale massacre to Assam. Videos and pictures in the message illustrated some tragic scene victims. As a result the message intensified panic of the public. At last more than 300,000 people left for a safe place. At last, the government had to inhabit mass texting.

In above incidents, the tapping phone is one kind of social event happened in our society but mapped on the web. By the mapping, social events spread, evolve and mutate in the web along with interaction with real world. And we call such events as social events mapped on web. The latter incident is caused by message on web and impact on real world. In other words, this kind of event happened in virtual world but evolve with human interference. We call such events as web sentiment events. All of these two kinds of events are called web event. Some web events have much bad influence on society. To avoid these bad influences, it is necessary to monitor and predict the evaluative tendency of web events. Therefore, how to collect and organize web events in the intelligent and automatic way, and how to track and measure dynamic evolution of web events are becoming an important subject in the field of information processing.

The evolution is a basic feature of web events and is also a part of studies on Topic Detecting and Tracking (TDT) [5-7]. Traditional TDT involves detecting unknown events, gathering and segmenting information, detecting when the event first reported, detecting follow-up reports of events and tracking events' tendency. Generally, TDT technology attempts to detect unknown web events and make related news pages clustered. Although TDT tracks development of web events, it does not measure the dynamic evolution process of web events. So we cannot have a global and clear

understanding of web events. In this paper, outbreak power based algorithm is proposed to measure the evolution process of web events in order to help people make decisions.

To calculate outbreak power of web events, it is necessary to get the temporal features of web events evolution firstly. Secondly, an iterative algorithm is put forward to get temporal features together in order to measure evolution course of web events. In addition, according to outbreak power in evolution course of web events, web events are classified into three classes: emergent event, popular event, general event. The detailed definitions of each class will be given later chapters.

At last, a fuzzy based algorithm for class discrimination of web events is introduced to verify that outbreak power can correctly describe the evolution process of web events. By measuring evolution process of web events, we discover features of each class events, which can be considered as prior knowledge to discriminate the class of web events.

The main contributions of this paper are:

(1) Definitions of web event classes are given;

(2) Outbreak power is introduced to estimate web event;

(3) Fuzzy theory based algorithm is presented to build prior knowledge for class discrimination;

(4) Fuzzy based algorithm for class discrimination is proposed.

The rest of the paper is organized as follow. Section 2 introduces temporal feature set in evolution process of web events. Section 3 introduces outbreak power of web events and discusses the method of its calculation in detail. Section 4 gives definitions of web events classes and gives some features which can be used in class discrimination of web events. Fuzzy cognition based algorithm for type discrimination is proposed in section 5. Conclusion is given in the last section.

2 Temporal Features and Basic Definitions of Web Events

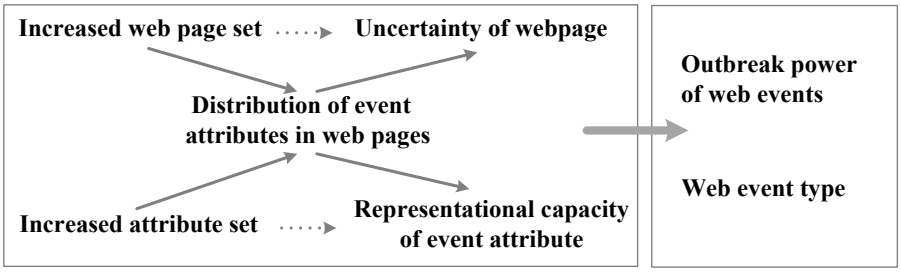

Figure 1. Framework of calculation of outbreak power of web events

This section focuses on the acquisition of temporal features of web events. Firstly we give the expressions of symbol and meaning of the expressions, as shown in Table 1. Framework of this is displayed in Fig.1. All temporal features cooperate on getting the result of outbreak power and web event types, and more details will be discussed later. The input and output of the algorithm is described as follows:

Input: a web event *e*, a set of temporal features (e.g., web pages, event attributes), event starting timestamp t_s, and current timestamp t_e.

Output: a set of evolution measurement of web event e, $op(t_s, t_e) = \{op_{t_s}, op_{t_{s+1}}, ..., op_{t_e}\}$, where op_{t_i} is the outbreak power of web event at time t_i.

Table 1: Symbol reference table

e	Web event
φ	Web page
$c(\varphi)$	The uncertainty of web pages φ describing the event e
$\Delta\varphi(t_i, t_j)$	Increased web page set during time t_i to time t_j
$w\phi(k)$	Web pages set of offering event attributes k
k	Attribute of web event
$\Delta k(t_i, t_j)$	Increased attribute set during time t_i to time t_j
$er(k)$	Representational capabilities of attribute to an event
$\psi(t_i, t_j)$	distribution of attributes in web pages
$op(t_i, t_j)$	Outbreak power of web events during time t_i to time t_j

2.1 Temporal Features of Web Events

Before we introduce outbreak power of event, some features and definitions are declared as follows.

Temporal feature 1: **Increased web page set $\Delta\varphi(t_i, t_j)$ of web event**

For a web event e, the increased web page set of web event $\Delta\varphi(t_i, t_j) = \{\varphi_i, \varphi_{i+1}, ..., \varphi_j\}$ is all the new published web pages from time t_i to t_j. And there is no intersection between increased web page sets, namely $\Delta\varphi(t_s, t_i) \cap \Delta\varphi(t_i, t_j) = \varnothing$.

Temporal feature 2: Increased attribute set $\Delta k(t_i, t_j)$ of web event

Increased attribute set $\Delta k(t_i, t_j) = \{k_i, k_{i+1}, ..., k_j\}$ of web event from time t_i to t_j is a cluster of attributes extracted from increased web page $\Delta\varphi(t_i, t_j)$. And an increased webpage φ_i can be represented by a vector, denoted as:

$$\varphi_i = \{w_{i1}, w_{i2}, ..., w_{im}\} \tag{1}$$

where $w_{ij} = (1 + \log tf(k_j)) * \log(1 + n / df(k_j))$. $tf(k_j)$ means the term frequency of attribute k_j in webpage φ_i, $df(k_j)$ means the webpage frequency of attribute k_j in $\Delta\varphi(t_i, t_j)$.

After the increased web pages $\Delta\varphi(t_i,t_j)$ and the increased attributes $\Delta k(t_i,t_j)$ are obtained, we also can get distribution of attributes in web pages by statistical method.

Temporal feature 3: Distribution of event attributes in web pages, $\psi(t_i,t_j)$

For a web event *e*, every webpage of $\Delta\varphi(t_i,t_j)$ can be denoted by attributes of $\Delta k(t_i,t_j)$, as shown in temporal feature 2. So, distribution of attributes in web pages can be denoted by a matrix as following:

$$\psi(t_i,t_j)=\begin{bmatrix} w_{11} & \cdots & w_{1m} \\ \cdots & \cdots & \cdots \\ w_{n1} & \cdots & w_{nm} \end{bmatrix} \tag{2}$$

where *n* is the number of webpages and *m* is the number of attributes.

Fig.2 shows the complete process of data source acquisition and the relationships of increased web pages of events $\Delta\varphi(t_i,t_j)$, increased attributes of events $\Delta k(t_i,t_j)$, and distribution of attributes in web pages $\psi(t_i,t_j)$. Every temporal feature describes web events in different aspects. Such as increased web pages describes physical characteristics of web events, increased attributes describes semantic characteristics of web events. By the further study on temporal features, we find evolution process of web event can be influenced by temporal features. So an iterative algorithm, which integrates three temporal features, is proposed to calculate the outbreak power and measure the evolution course of web events.

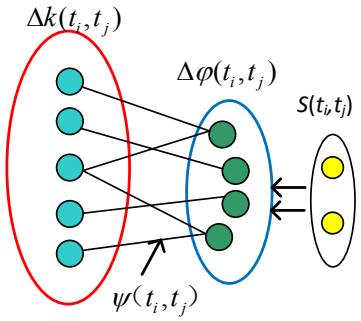

Figure 2. Seed $S(t_i,t_j)$ (will be discussed in experiment section), increased web pages of events $\Delta\varphi(t_i,t_j)$, increased attributes of events $\Delta k(t_i,t_j)$, and distribution of attributes in web.

However, we have just considered the text-oriented information; in fact user-oriented information of web events is also important, such as Google Trends, which calculates the hotness of web events by the number of searches. User-oriented information is generally not open to public so that it is difficult to obtain. Furthermore, user-oriented information is typically lagged. That is because user may respond to a web event, which has taken place for a period of time. Therefore, here we take only data-oriented information into account.

Text-oriented data that we use is about text information of web events. In addition, pictures and videos are also the information that describes web events. And these kinds of information can be obtained by search engine, but it is hard to deal with this information directly. So we consider for only text information of web events.

In this section, we first introduce the definitions about outbreak power of web events. Outbreak power describes the evolution process of web events and also reflects the changes of temporal features in different time intervals. Furthermore, we proposed an iterative algorithm that integrates temporal features to measure the evolution course of web events.

2.2 Basic Definitions About Outbreak Power of Web Events

Outbreak power can directly show the changes of temporal features and help people clearly know the evolution of web events without reading a large number of relevant web pages. To make it clear, following definitions are put forward.

Definition 1. Outbreak power of web events, $op(t_i, t_j)$

For a web event e, the outbreak power $op(t_i, t_j)$ means the emergency degree of corresponding web event during time t_i to t_j. Outbreak power of web event is influenced by two elements denoted as $op(t_i, t_j) =< er(k), c(\varphi) >$. $er(k)$ is the attribute representational capacity and $c(\varphi)$ is the webpage uncertainty, two of which will be discussed later.

If webpage of web events is quite uncertain and all the attributes have strong representational capacity then the development direction of web event may diverse. In this situation, the outbreak power of web event is high.

Definition 2. Representational capacity of event attribute, $er(k)$

Representational capacity of event attribute is the ability of attribute describing an event, donated as $er(k)$.

For instance, if all the information of web event is described by a certain group of attribute, then the representational capacity of this web event is high, and as the result the outbreak power of web event should be low. On the contrary, the outbreak power of web event should high.

Definition 3. Uncertainty of webpage, $c(\varphi)$

The uncertainty of webpage is the ability of webpages in describing corresponding web event. It is related to general representational capacity of event attribute.

For instance, if the attributes of webpages about certain web event diverse a lot, then representational capacity of these attributes is weak and outbreak power of this web event should be high.

Corollary 1. Suppose there exist some web pages and attributes discussing the same event, and the distribution of attributes in web pages approaches to one-to-one mapping, then it means the discussions of web event are different so that the outbreak power of web event is high.

According to **Corollary 1**, if each attribute is only provided by one webpage and each webpage provide only one attribute, the degree of similarity between each page is 0. So the contents of all web pages are different and discussions of web event are different, which may lead to further deteriorate of event. In contrast, we have come to **Corollary 2**.

Corollary 2. Suppose there are limited web pages and attributes discussing an event, and the distribution of attributes in web pages approaches to a complete graph, then it means the discussion opinions of web event are almost the same so that the outbreak power of web event is low. That is:

$$(\forall w_{nm} \in \psi(t_i, t_j) \to w_{nm} \neq 0) \to op(t_i, t_j)_{\min}$$

According to **Corollary 2**, the distribution of attributes in web pages approaches to a complete graph, that is, events attributes exist in each webpage. And the degree of similarity between each page is 1, which means all pages are reproduced from a webpage and discussion opinions of web event are same.

From above corollaries, we can know the most influences to outbreak power are: increased web pages $\left|\Delta\varphi(t_i, t_j)\right|$, increased attributes $\left|\Delta k(t_i, t_j)\right|$, and distribution $\psi(t_i, t_j)$ of attributes in web pages.

So according to representational capacity of event attributes, we can infer the uncertainty of a webpage. And we also calculate the representational capacity of event attributes by the uncertainty of a webpage. Fig.3 is composed of three parts: increased web pages $\left|\Delta\varphi(t_i, t_j)\right|$, increased attributes $\left|\Delta k(t_i, t_j)\right|$, and distribution $\psi(t_i, t_j)$ of attributes in web pages.

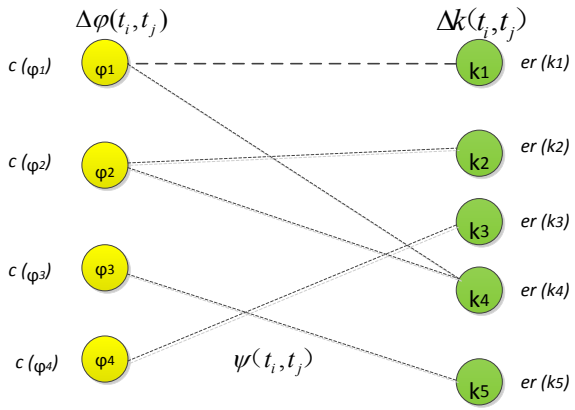

Figure 3 Iteration computation of the uncertainty of increased web pages and the representational capacity of increased attributes through distribution of attributes in web pages.

According to Fig 3, we get two corollaries:

Corollary 3. For a webpage φ_i, if it provides most of the event attributes with strong representational capacity $er(k)$, this webpage has low uncertainty φ_i.

Corollary 4. For an event attribute, if it is provided by a webpage with low uncertainty $c(\varphi_i)$, this event attribute has strong representational capacity $er(k)$.

Corollary 3 and 4 are called "iterative Corollary ", by the two corollaries, we find the uncertainty of web pages and representational capacity of event attributes have interactions, so we proposed an iterative algorithm to calculate them. And the iterative algorithm uses increased web pages, increased event attributes, and distribution of attributes in web pages as variables.

Corollary 5. The higher the uncertainty of overall increased web pages $\left|\Delta\varphi(t_i,t_j)\right|$ about an event, the higher the outbreak power of this event.

As we know, in a certain time interval $[t_i,t_j]$ and for an event, if the uncertainty of overall web pages is high, the semantic of this event changes a lot and event may breakout. If the uncertainty of overall web pages is low, this event stabilizes and is less likely to breakout.

3 Outbreak Power Calculation of Web Events

In this chapter, we mainly introduce the calculation of outbreak power of web events. According Corollary 2, the uncertainty of a webpage describes representational capability of event attributes provided by webpage itself. For a webpage φ, its uncertainty $c(\varphi)$ can be obtained by calculating the representational capability of event attributes it provides:

$$c(\varphi)=1-\frac{\sum_{k\in Kw(\varphi)} er(k)}{|\,Kw(\varphi)\,|} \tag{3}$$

where $Kw(\varphi)$ is the attribute set provided by webpage φ.

According to Eq.3, the uncertainty of web pages is related to the representational capability of event attributes provided by web pages. If the representational capability of event attributes is strong, the uncertainty of web pages is low. If the representational capability of event attributes is weak, the uncertainty of web pages is high.

According to Corollary 3, the uncertainty of a webpage depends on representational capability of event attributes provided by this webpage. Corollary 3 and Corollary 4 reflect the interrelationship between the uncertainty of web pages and representational capability of event attributes.

The calculation of the representational capability of event attributes $er(k)$ is complex, according to [11], we use probability function to calculate the representational capability of event attributes:

$$er(k)=1-\prod_{\varphi\in Wp(k)} c(\varphi) \tag{4}$$

where $Wp(k)$ is the webpage set which provides the attributes *k*.

According to Eq. 4, if the uncertainty of web pages is low, the representational capability of event attributes is weak; on the contrary, if the uncertainty of web pages is high, the representational capability of event attributes is strong.

The outbreak power of web events $op(t_i,t_j)$ in a certain time interval $[t_i,t_j]$ can be calculated by the formula:

$$op(t_i,t_j)=\sum_{q=1}^{n}c(\varphi_q) \tag{5}$$

where *n* denotes the increased web pages in time interval $[t_i,t_j]$.

According to Eq.5, the outbreak power of web events in time interval $[t_i,t_j]$ is related to the uncertainty of corresponding web pages. If the uncertainty of web pages is high, the outbreak power of web events is high; if the uncertainty of web pages is low, the outbreak power of web events is low.

Since the calculation of the Eq. 3 is based on Fig 3, increased web pages of events $\Delta\varphi(t_i,t_j)$, the influences of increased attributes of events $\Delta K(t_i,t_j)$, and distribution of attributes in web pages $\psi(t_i,t_j)$ are all taken into consideration.

We propose algorithm1 to calculate the outbreak power of web events are as Table 3.

Table 3 Steps of computing outbreak power

Algorithm 1: Computing Outbreak Power

Input: The set of pages $\Delta\varphi(t_i,t_j)$, the set of keyword $\Delta K(t_i,t_j)$, the distribution of keywords among pages $\psi(t_i,t_j)$.

Output: The outbreak power $op(t_i,t_j)$.

for each $\varphi\in\varphi(t_i,t_j)$

$c(\varphi)\leftarrow c_0$

$\vec{T}=\{c_0(\varphi_1),c_0(\varphi_2),...,c_0(\varphi_n)\}$

repeat

$\overrightarrow{er(k)}\leftarrow\overrightarrow{c(\varphi)}$

$\overrightarrow{c'(\varphi)}\leftarrow\overrightarrow{er(k)}$

$\vec{T}'=\{c'(\varphi_1),c'(\varphi_2),...,c'(\varphi_n)\}$

Until cosine similarity of $\vec{T}$ and $\vec{T}'$ is greater than β.

Compute $op(t_i,t_j)$ by Equation 5

From above steps of algorithm, the uncertainty of web pages and the representational capability of event attributes are obtained by iterative calculation. According to [11,23], just like other iterative algorithm, we set initial uncertainties of all web pages to α as the initial state value. Similar to the famous Page Rank algorithm of Google, Page Rank algorithm assumes the rank of all web pages are the same initially and first rank can be calculated based on the initial rank, and then the second rank can also be calculated based on the first rank. Larry Page also theoretically proves that this algorithm guarantee the result of Page Rank converges to the true value regardless of how to select the initial values. In addition, Page Rank algorithm is completely without any manual intervention and it is

brilliant that this algorithm treats the entire internet as a whole, which is in line with the view of system theory.

4 Web Event Types

In the previous section, we have gotten the time series data of outbreak power. In order to verify that the outbreak power can reflect the evolution of web events objectively, fuzzy based algorithm for type discrimination of web events is proposed. Provided the result of discrimination in line with objective facts and knowledge of human cognition, we believe that outbreak power can better describe the evolution of web events and the algorithm for the calculation of outbreak power is reasonable.

By analysing the time series data of outbreak power, we can extract some important features that describe the event itself. For example, outbreak power and fluctuation power of web events are the two kinds of features. And by mining features pattern, membership function of each type event can be established, which can be the prior knowledge to help us study web events. As each type event has its own law and features, we can discriminate the type of web events by means of prior knowledge. Next we will introduce fuzzy based algorithm framework use outbreak power of web events to discriminate the type of web events. For the above discussion, we define web event first.

Web event *e* is a tuple $(L_e, op(L_e))$, where L_e is a time interval that from starting timestamp t_s to ending timestamp t_e (the time interval of life course of web event is L_e), the outbreak power of web event e, presents the changes of semantic features of web event in time interval L_e.

A great number of web events are provided on web every day, internet users have to waste a lot of time and patience to search the interesting topics, so it has become a problem that internet users how to find events they are interested in time. Therefore, it is necessary to classify the web events into different types and recommend web events to users by types. Web events classification is based on TDT system. Current studies mainly focus on classification of web text, but studies on type discrimination of web events are rare. Therefore, we will give the definition of web events type.

Definition 4. The type of web events, ε

Let $E = \{e_1, e_2, ..., e_s\}$ denote the set of web events. Hypothesis space of web events type refers to the set of all type of web events on web, denoted as $\varepsilon = \{\varepsilon_1, \varepsilon_2, ..., \varepsilon_n\}$, $\varepsilon_i \subseteq E$, where ε_i means a certain type of web events, *n* is the size of the hypothesis space.

According to the observation of actual data set and cognitive knowledge, and by taking account nature, severity, controllability and influence of web events, web events are classified into three types, that is, $\varepsilon = \{\varepsilon_1, \varepsilon_2, \varepsilon_3\}$, where ε_1 is emergency event, ε_2 is hot events, ε_3 is general event.

Suppose $TS = \{op_1, op_2, ..., op_n\}$ denotes the time series data of outbreak power of web events in life course L_e, where op_i is the outbreak power at time t_i, *n* is the length of time series. For a web event, it can be described with a successive time interval, that is, $TS = \{TS_1, TS_2, ..., TS_{k-1}, TS_k\}$, where k is the number of time interval. A time interval is denoted as $TS_k = \{op_a, op_{a+1}, ..., op_b\}$, where subscript of *op* represents corresponding timestamp and t_a is the starting timestamp and t_b is the ending timestamp in time interval.

Through the segmentation of time series data of outbreak power, evolution course of a web event can be divided into different segments. For describing different segments of evolution course, the following temporal features are proposed:

(1) Average Outbreak Power, $op_{ave}(TS_k)$

$$op_{ave}(TS_k) = \frac{\sum_{i=t_a}^{t_b} op_i}{t_b - t_a} \tag{6}$$

where op_i is the outbreak power of web events at time t_i, $[t_i, t_j]$ is the time interval of event segment TS_k. Average outbreak power $op_{ave}(TS_k)$ reflects the level of outbreak power in segment TS_k.

(2) Fluctuation Power, $fp(TS_k)$

$$fp(TS_k) = \frac{Var}{op_{ave}(TS_k)},\ Var = \sqrt{\frac{1}{t_b - t_a}\sum_{i=t_a}^{t_b}(op_i - op_{ave}(TS_k))^2} \tag{7}$$

where op_i is the outbreak power of web events at time t_i, $[t_a, t_b]$ is the time interval of event segment TS_k. Fluctuation power $fp(TS_k)$ reflects the fluctuation degree in segment TS_k.

Obviously, the type of a web event will change with the evolution of this event. Therefore, in order to discriminate the present type of a web event, we must consider not only the historical evolution course of this event but also the current evolution state of event. Then synthesized outbreak power is proposed as follows:

$$op_{ave}(TS) = \sum_{i=1}^{k} w_i * op_{ave}(TS_i),\ \ \sum_{i=1}^{k} w_i,\ 0 < w_1 < w_2 < \ldots < w_{k-1} < w_k < 1$$

where TS_k is the *i-th* segment, w_i is the weight corresponding to TS_i, k is the number of segments. In addition, the recent segment has higher weight.

Similarly, formula to calculate the synthesized fluctuation power is as follows:

$$fp(TS) = \sum_{i=1}^{k} w_i * fp(TS_i),\ \ \sum_{i=1}^{k} w_i = 1,\ 0 < w_1 < w_2 < \ldots < w_{k-1} < w_k < 1$$

where TS_k is the *i-th* segment, w_i is the weight corresponding to TS_i, k is the number of segments. The recent segment has higher weight.

From the above two Equations, in order to calculate the synthesized fluctuation power and the synthesized outbreak power, the historical evolution course and the current evolution state must be taken into account.

After the discussion of web event, we give the definition of each type of web events:

Definition 5. Urgent event, ε_1

Urgent event is the event caused by major natural disasters, accidents, or social security that required to be responded to and deal with by government or social groups within a special time interval. Such

as "5.12 Wenchuan Earthquake", "9.11 terrorist attacks" and "7.23 Wenzhou motor car accident". This type event usually has features of sudden, complexity, destructive, persistent.

According to the outbreak power observation of emergent events in their life courses, when an event evolves into an emergent event, this event has high synthesized outbreak power $op_{ave}(TS)$ and synthesized fluctuation power $fp(TS)$, as shown in Fig 4(a), that is:

$$\varepsilon_1 = \{E' \mid E' \subseteq E \wedge (\forall e \in E') \rightarrow (op_{ave}(TS) > \delta_1 \,\&\&\, fp(TS) > \theta_1)\}$$

where $\delta_1, \delta_2, \theta_1, \theta_2$ are the threshold of synthesized outbreak power and synthesized fluctuation power, respectively.

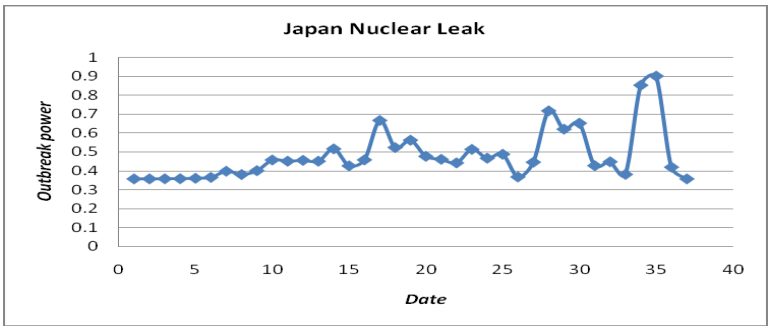

Figure4(a) Evolution course of outbreak power of event "Japan Nuclear Leak"

Definition 6. Hot event, ε_2

Hot event is the event that related to people daily life happening in real society or on web and people concerns it for a long term. Such as "Price regulation of house", "Food security".

According to the outbreak power observation of hot events in their life courses, we discover that hot events usually have moderated synthesized outbreak power $op_{ave}(TS)$ and synthesized fluctuation power $fp(TS)$, as shown in Fig 4(b), that is:

$$\varepsilon_2 = \{E' \mid E' \subseteq E \wedge (\forall e \in E') \rightarrow (\delta_2 < op_{ave}(TS) \le \delta_1 \,\&\&\, \theta_2 < fp(TS) \le \theta_1)\}$$

where $\delta_1, \delta_2, \theta_1, \theta_2$ are the threshold of synthesized outbreak power and synthesized fluctuation power, respectively.

Definition 7. General event, ε_3

General event is the event that happens in real society or on web and attracted less attention compared with hot event. Usually this type event is reported when it occurs, and then forgotten by people quickly. For example, "Super moon" and "Forbes Chinese rich list" are the typical general events.

According to the outbreak power observation of hot events in their life courses, we find that general events usually have low synthesized outbreak power $op_{ave}(TS)$ and synthesized fluctuation power $fp(TS)$, as shown in Fig 4(c), that is:

$$\varepsilon_3 = \{E' \mid E' \subseteq E \wedge (\forall e \in E') \rightarrow (op_{ave}(TS) < \delta_2 \,\&\&\, fp(TS) < \theta_2)\}$$

where δ_2, θ_2 are the threshold of synthesized outbreak power and synthesized fluctuation power, respectively.

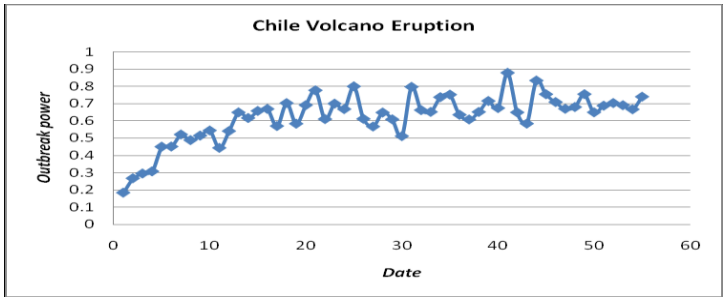

Figure 4(b) Evolution course of outbreak power of event "Chile Volcano Eruption"

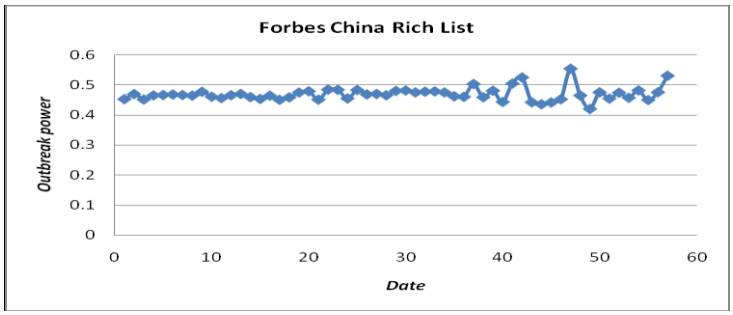

Figure 4(c) Evolution course of outbreak power of event "Forbes China Rich List"

5 Fuzzy Based Algorithm for Type Discrimination of Web Events

With the time changing, the emergent degree of web events changes dynamically. One event in different segments has different emergent degree, so for a web event, it may go through three states: general state, hot state, and emergent state. Few domestic and foreign scholars study on emergent level classification of web events in different segments, so that the lack of a prior knowledge of type discrimination of web events in different segments. Therefore, we study the changes of features and emergent degree of web events in evolution course, and we can obtain the relationship between outbreak power and fluctuation power. Then by studying these relationships, we extract features of different emergent degree, establish evolution model of web events, and construct the membership

model for type discrimination of web events as prior knowledge. Thereby to provide effective guidance for the type prediction of web event in later section.

In this section, we will introduce how to build a prior knowledge for type discrimination of web events. So based on outbreak power we propose a fuzzy based algorithm for type discrimination of web events.

5.1 Adaptive Segmentation Algorithm of Outbreak Power of Web Events

Herein, we first discuss how to divide a web event into different segments in its evolution course. In this paper, we adopt k segmentation method in [24], this method can find the optimal segment boundaries: $c_1, c_2, ..., c_{k-1}$, $0 < c_1 < c_2 < ... < c_{k-1} < N$, then divide the time series of outbreak power into k segments.

According to [24], it is the optimum segmentation of time series data only if each segment internal is homogeneous, that is, the event is in the same evolution state in one segment, and different segments represent different evolution stages in event evolution course. So we introduce segment cost function for segmentation and propose two hypotheses.

Single segment cost function $cost_F(TS_k)$ is the function of data point values in segment and length of segment $l = t_b - t_a$ (the number of data point in segment).

$$cost_F(TS_k) = F(op_i, l), \quad op_i \in TS_k \tag{8}$$

For a web event, segment cost function $cost_F(TS_k)$ represents the fluctuation degree of outbreak power of this event in a certain time interval, that is, changes in frequency and amplitude of this event within a period of time.

Total cost of K segments is the sum of each segment $TS_1, TS_2, ..., TS_k$:

$$cost_F(TS) = cost_F(TS_1 TS_2 ... TS_k) = \sum_{i=1}^{k} cost_F(TS_i) \tag{9}$$

Total cost function $cost_F(TS_1 TS_2 ... TS_k)$ represents the fluctuation degree of outbreak power of an event in its life course L_e.

The above functions are abstract functions. Specifically, we use the variance of the segmented data to measure the fluctuation of outbreak power, the formula is as follows:

$$cost_F(TS_k) = \frac{1}{l}\sum_{t=t_a}^{t_b} op_i^2 - \left(\frac{1}{l}\sum_{t=t_a}^{t_b} op_i\right)^2 \tag{10}$$

In its evolution course, web event typically go through different stages, event in one stage is in a same state, and event in different stages is in different states. So we believe that fluctuation degree of outbreak power is low in one stage and fluctuation degree of outbreak power is low between different stages. Therefore, segmentation of time series data can be transformed to find optimal segment boundaries to minimize the total cost, that is, $(\exists\{c_1, c_2, ..., c_{k-1}\} \to cost_F(TS_1 TS_2 ... TS_k)_{\min}) \to TS_F^{opt}(TS_i, k)$

5.2 Fuzzy Based Membership Function

In previous sections, temporal features are proposed, for example, average outbreak power $op_{ave}(TS)$, fluctuation power $fp(TS)$. Here, we use a set of feature vectors to represent an event. For a web event e, we can get a feature vector by calculating the time series data of outbreak power of this web event, denoted as $\overrightarrow{T_d(e)} = \{op_{ave}(TS), fp(TS)\}$.

Obviously, the features of web event can be obtained directly from the time series data of outbreak power, and the different types of events should have different feature patterns. Therefore, we need to establish the connection and mapping from temporal features of web events to hypothesis space of event types. According to [13-17], fuzzy math theory is used to discriminate the type of web events. We introduce membership grade to measure degree of temporal feature belonging to a particular type ε_i of web event, denoted as $\xi_i(op_{ave})$ and $\xi_i(fp)$. Through the membership function, we have established a relationship between the temporal features and event types.

Membership function reflects the distribution of temporal feature belonging to event type ε_i when temporal feature get different values.

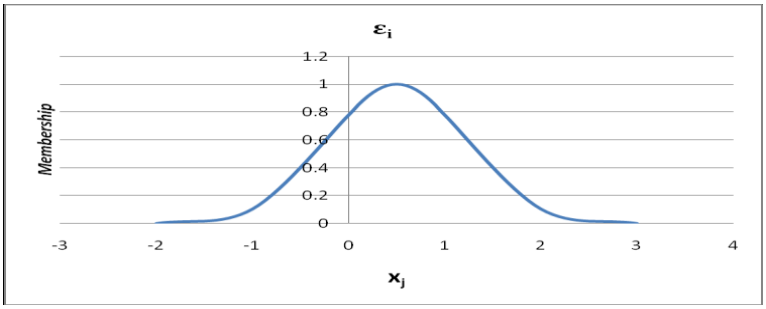

Figure 5 Membership distribution of temporal feature belonging to event type ε_i

For a known web events data set which is a manual annotated training data set, membership frequency of temporal feature x_j belonging to a certain event type ε_i for different values can be obtained by statistical, that is the member grade of temporal feature x_j belonging to event type ε_i. Further, summarizing the membership grade of each temporal, we can get membership function vector for a certain type of web event, that is:

$$\vec{\varepsilon_i} = \{\xi_i(op_{ave}), \xi_i(fp)\} \tag{11}$$

Obviously, each type of web events has its own laws, and in this paper, membership function, which is used to describe the models of each type event, reflects the event its own laws.

5.3 Fuzzy Based Algorithm for Type Discrimination of Web Events

After the establishment of prior knowledge of type discrimination, we can use this prior knowledge discriminate the type of unknown web events.

For a web event e, its feature vector denoted as $\overrightarrow{T_d(e)} = \{op_{ave}(TS), fp(TS)\}$, so its membership vector of belonging to event type i is as follows:

$$\overrightarrow{\varepsilon_i(e)} = \{\xi_i(op_{ave}), \xi_i(fp)\} \tag{12}$$

where $\xi_{ij}(x_j)$ is the membership of the *j-th* feature component belonging to *i-th* type event.

And then, we can get the membership of this event e belonging to *i-th* type event:

$$\varepsilon_i(e) = \xi_i(op_{ave}) * \alpha_1 + \xi_i(fp) * \alpha_2, \quad \alpha_1 + \alpha_2 = 1 \tag{13}$$

where α_j is the weighting factor of *j-th* feature component. Taking into account the role played by each feature component in type discrimination may be different, so give different feature components different weights. Assuming the importance of each feature component is the same, that is:

$$\varepsilon_i(e) = \xi_i(op_{ave}) * 0.5 + \xi_i(fp) * 0.5 \tag{14}$$

Next, based on the principle of maximum membership in fuzzy math, the corollary 6 are proposed as the basis for type discrimination of web events.

Corollary 6. For a certain type of web event ε_i, if web event e has the largest membership $\varepsilon_i(e)$ for event type ε_i, this event e belongs to this event type ε_i, that is:

$$(\exists i \in \{1,2,...,n\}) \rightarrow \varepsilon_i(e) = \max\{\varepsilon_1(e), \varepsilon_2(e), ..., \varepsilon_3(e)\}) \rightarrow e \in \varepsilon_i$$

6 Experiments and Analysis

6.1 Experiment Data Set

In this paper, we set seeds (for example, “Japan”, “earthquake”, and “tsunami ”) of a web event from some news websites (such as “baidu.news”). Then search these seeds as keywords making use of search engines such as Google, and a number of corresponding web pages are collected as shown in Table 2. And from web pages of web events, we can get increased web pages of web events, increased attributes of web events, and distribution of attributes in web pages, as shown in Tab 3.

In this paper, the web events involved in the experiment are derived from *Baidu* news sites. *Baidu* provides daily popular events which are searched by a large number of users. In the experiment, we selected approximately 100 web events about 900,000 pages were selected as the set of experimental data These events cover topics of political, accident, disaster, terrorist attacks in various fields. Table 2 show the statistical results of experimental data set in this article in detail. The starting timestamp chosen for these web events was according to [12], and the ending timestamp is the time when we get the events from web pages, and the average length of sampling time for each web event was about 30-40 days. In addition to determining the life course of web events, we also got their semantic features, including the seed set, web pages set, and key worlds set. Seed set of web events was extracted from *Baidu*. *Baidu* provides popular events and also provide hot search words to help users search for web events.

Table 2 The details of dataset used to period detection (100 events)

Feature	Value
Average number of seeds per event	2
Average number of Webpages per event	5556
Average number of event attributes per event	16856
Average number of days per event	40
Average number of Webpages per day	146
Average number of event attributes per day	469

For example, Japan suffered catastrophic earthquake and tsunami in 2011, many internet users wanted to search related web pages to understand the situation, so *Baidu* provided a set of key word of this event (e.g., Japan, earthquake, tsunami). After obtained the seed set, we used seed set as search words to search web events and got the related web pages. The detail steps in our experiment are as follows:

(1) Obtain seed set $S(t_i,t_j)$ of web events from Baidu or other news sites, (e.g., Japan, earthquake, tsunami).

(2) Use seed set $S(t_i,t_j)$ as search words to search web events and got the related web pages set $\varphi(t_i,t_j)$.

(3) Determine the starting timestamp t_s of web events from web pages set $\varphi(t_i,t_j)$, and ending timestamp t_e is determined according to the time we crawled the events from web pages.

(4) Obtain the daily time series source data from web pages set $op(t_i,t_j)$, (i.e., increased web pages $|\Delta\varphi(t_i,t_j)|$, increased attributes of events $|\Delta K(t_i,t_j)|$, distribution of attributes in web pages $\psi(t_i,t_j)$.

(5) Perform the above steps for different information sources (i.e., blog, BBS), respectively.

6.2 Experiments on Outbreak Power of Web Events

A. Instance verification for outbreak power of web events

For the result of algorithm, it describes the emergent degree of web events and it is called outbreak power. In this paper, "day" is the minimum time granularity. Source data of temporal features of web events is collected from different news sites daily, algorithm 1 calculate the daily outbreak power of web events based on these source data, and then time series data of outbreak power of web events in a certain time interval are obtained, as shown in Fig 6.

The outbreak power of web events, which is calculated by algorithm 1, combines the increased webpages, increased attributes of events, and distribution of attributes in webpages. The

algorithm 1considers the physical attributes of web events, semantic content, and distribution of web events on web. So the outbreak power we get can comprehensively describe the evolution course of web events.

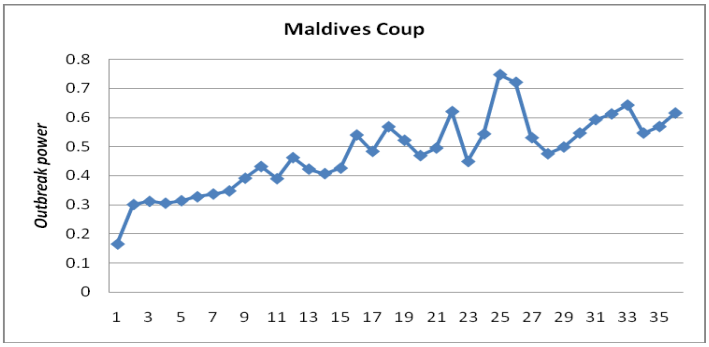

Figure 6 Temporal data of outbreak power of event "Maldives Coup"

B. Outbreak power based correctness verification for type discrimination of web events

Herein, 100 web events were selected as the experimental object; the detailed statistical results are shown in Table 2. And 60 web events among experimental object as training set to establish prior knowledge of web events, and the remaining 40 web events were as test set of type discrimination.

In experiment, we first trained 60 web events in training set, annotated the web events according to their emergent degree, so these 60 web events were labelled as emergent event, hot event or general event. By statistics on the training set, we calculated the membership frequency of temporal features belonging to each type when temporal features took different values, and combined with prior knowledge of our cognition on web events, we got the membership distribution of each temporal feature belonging to different types of web events. Here, Fig 7 shows the membership distribution of average outbreak power belonging to different types of web events based on the cases used in the experiment.

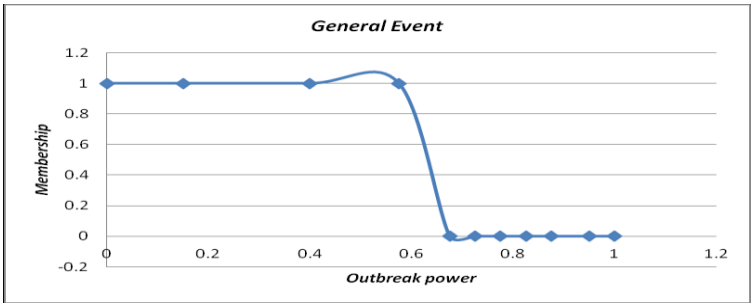

Fig 7(a) Membership function distribution of outbreak power of general events based on the cases used in experiment

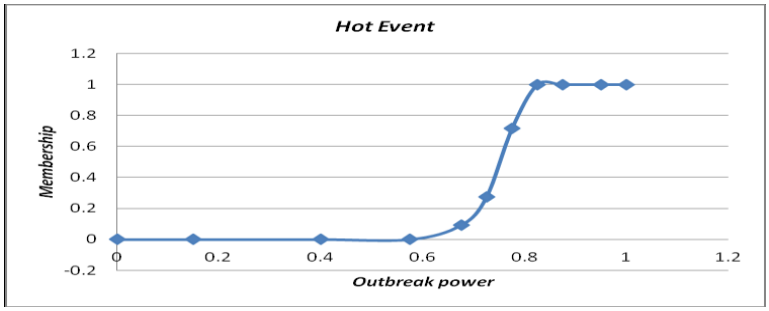

Fig 7(b) Membership function distribution of outbreak power of hot events based on the cases used in experiment

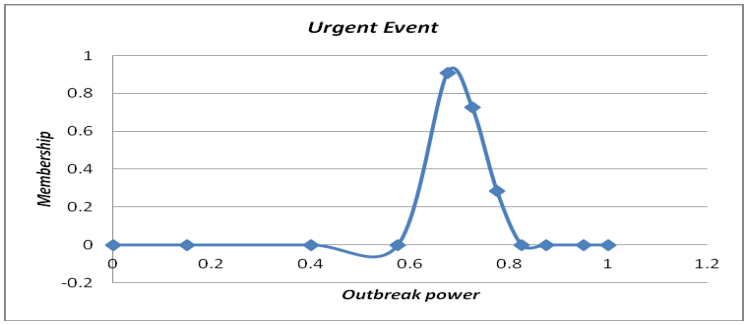

Fig 7(c) Membership function distribution of outbreak power of urgent events based on the cases used in experiment

And Fig 8 shows membership distribution of fluctuation power belonging to different types of web events based on the cases used in the experiment.

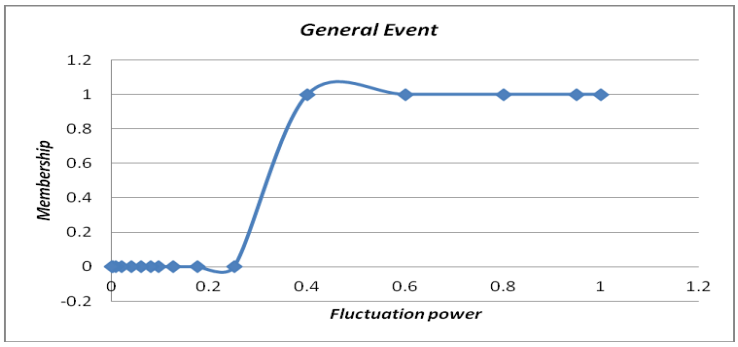

Fig 8(a) Membership function distribution of fluctuation power of urgent events based on the cases used in experiment

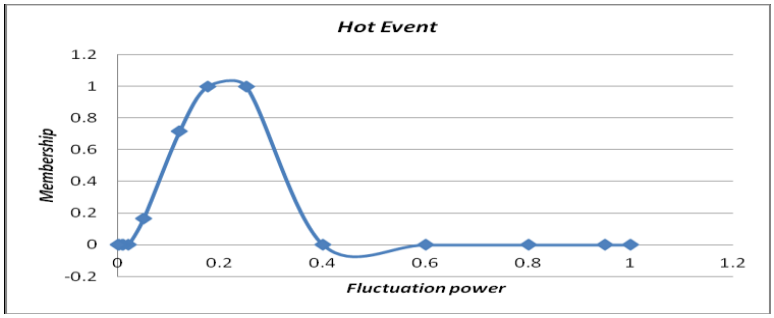

Fig 8(b) Membership function distribution of fluctuation power of hot events based on the cases used in experiment

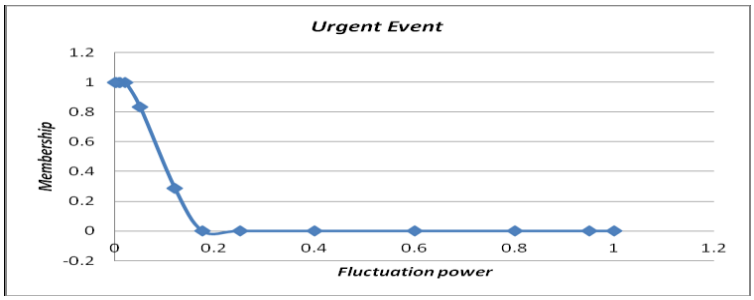

Fig 8(c) Membership function distribution of fluctuation power of urgent events based on the cases used in experiment

By fitting the above discussed membership distributions, we got the membership function of each type of web events. Eq.(15) and Eq.(16) shows the membership function of outbreak power and fluctuation power belonging to each type of web events based on the cases used in the experiment, respectively. To this end, prior knowledge of web events was established. Through these membership functions, we can easily found that outbreak power of web event declines in the order of hot event, urgent event and general event, also fluctuation of event declines in the order of general event, hot event and urgent event.

$$\xi_{11}(x_1)=\begin{cases}1, & x_1 \le 0.575\\ \dfrac{0.675-x_1}{0.675-0.575}, & 0.575 < x_1 \le 0.675\\ 0, & x_1 > 0.675\end{cases}$$

$$\xi_{21}(x_1)=\begin{cases}0, & x_1 \le 0.575\\ 20x_1^2-24x_1+7.2, & 0.575 < x_1 \le 0.825\\ 1, & x_1 > 0.825\end{cases} \qquad (15)$$

$$\xi_{31}(x_1)=\begin{cases}0, x_1 \le 0.575\\ -52x_1^2+73x_1-25, 0.575 < x_1 \le 0.825\\ 0, x_1 > 0.825\end{cases}$$

$$\xi_{12}(x_2)=\begin{cases}0, & x_2 \le 0.25\\ \dfrac{x_2-0.25}{0.4-0.25}, & 0.25 < x_2 \le 0.4\\ 1, & x_2 > 0.4\end{cases}$$

$$\xi_{22}(x_2)=\begin{cases}0, & x_2 \le 0.02\\ \dfrac{x_2-0.02}{0.175-0.02}, & 0.02 < x_2 \le 0.175\\ 1, & 0.175 < x_2 \le 0.25\\ \dfrac{0.4-x_2}{0.4-0.25}, & 0.25 < x_2 \le 0.4\\ 0, & x_2 > 0.4\end{cases} \tag{16}$$

$$\xi_{32}(x_2)=\begin{cases}1, & x_2 \le 0.02\\ \dfrac{0.175-x_2}{0.175-0.02}, & 0.02 < x_2 \le 0.175\\ 0, & x_2 > 0.175\end{cases}$$

Next, we discriminated the type of 40 web events with obtained prior knowledge. For each web event e, we got a set of feature vectors $\overrightarrow{T_d(e)}=\{x_1,x_2,...,x_m\}$ according to Eq.(6) and Eq.(7); and then, by means of membership function Eq.(10) and Eq.(12), membership $\varepsilon_i(e)$ of web event e belonging to each type of web events can be obtained. At last, according to corollary 6, we can discriminate web event e belong to which event type exactly.

Table 4 precision and recall rate of each type of web events of the cases in experiment

	Type of web events			total
	Urgent event	*Hot event*	*General event*	
Recall	95.2%	88.8%	90%	91%
Precision	90.9%	88.8%	85%	88%

A major evaluation of results is that correct rate of type discrimination in test set. Obviously, the higher the correct rate, the better. In order to verify that temporal features can describe the evolution course of web events for each web event in test set, it was manually annotated to justify whether the algorithm for type discrimination is objective and correct, which indirectly proof the correctness of the construct of outbreak power. Such as the event “Japan Nuclear Leak” is discriminated as hot event by algorithm, but if label members believe that this event should be labelled as emergent event, then we think that the event discrimination failed. In addition, each label member were arranged independently complete the evaluation in order to ensure that the reliability and effectiveness of the experimental results. Before the label member evaluated the experimental results, we provided the abstract description of each type of web events. Such as, we gave them some news samples and their concepts. The training stopped until label members were familiar with each type of web events. At last, all the evaluation results of each label member were aggregated together, and to reach a consensus. Table 4 shows statistical results of web events type discrimination. The overall

accurate rate of type discrimination reaches 88% and recall rate is 91%, which shows the good performance of our algorithm.

7 Conclusions and Future Work

The main contributions of this paper are as follows:

1) We propose outbreak power of web events. For any event, by analysing its information on web we can obtain a series of temporal features. Furthermore, outbreak power of web event is mined, which helps us get a clear understanding of evolution course of web events. All the temporal features, outbreak power and fluctuation of outbreak power are used in discriminating web events. And the experiment on real dataset promises a good performance.

2) We propose an algorithm framework for analysing web events. There are four steps in this framework. First, the information of web events are collected during their life courses. Second, the increased webpages, the increased event attributes, distribution of attributes in webpages, and the relationships of attributes are embedded into the calculation of outbreak power of web events. Third, prior knowledge of each event type is mined from outbreak power. Forth, a fuzzy based algorithm is introduced to discriminate the type of web events. Through these steps, we get a clear understanding of web events evolution during their life course.

3) We present the definitions of web event type. Web events are classified into three types: urgent event, hot event and general event. By analysing each type of web events, we can get general events have low outbreak power and high fluctuation, while urgent events and hot events have high outbreak power and low fluctuation, which is the prior knowledge of web events based on the cases in experiment. By employing these prior knowledge, we also can discriminate the type of web events and do further analysis and research on this basis.

However, there are still some shortcomings in our work:

1) Due to limited collection and unavoidable noise of our data set in the experiment, there are some negative impacts on the results of type discrimination of web events to some degree. All the results are based on the cases in our experiments. Future study is needed to decide whether the rules are suitable for other cases, even though the framework of getting outbreak power of events is general.

2) In this paper, we only focus on the evolution of semantic features during the life course of web events. Obviously, the evolution of web events is also affected by other factors. Such as, uncertainty of source websites, uncertainty of web events, and source difference of web events, etc. These factors also determine the web events in different aspects.

- **acknowledgment**

The research work in this paper was supported by the National Science Foundation of China (grant no. 61471232 and 91024012).

References

1. http://definitions.uslegal.com/e/emergency-event/, 2015.
2. D. Haddow, A. Bullock, and P. Coppola. Introduction to Emergency Management, 2010.

3. http://en.wikipedia.org/wiki/News_of_the_World_phone_tapping_scandal, 2015.
4. http://news.163.com/12/0820/08/89BC87R700014JB6.html, 2015
5. C. Yang, X. Shi, and C. Wei. Discovering Event Evolution Graphs from News Corpora. *IEEE Trans. on Systems, Man and Cybernetics—Part A:* 39(4):850-863, 2009.
6. Juha Makkonen. 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, 2003.
7. J. Allan, G. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic Detection and Tracking Pilot Study Final Report. In *Proceedings of the Broadcast News Transcription and Understanding Workshop*, 1998.
8. Wang, C. & Wang, S. (2000). Supporting content-based searches on time Series via approximation. Proceedings of the 12th International Conference on Scientific and Statistical Database Management.
9. Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., & Allan, J. (2000). Mining of Concurent Text and Time Series. Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining. pp. 37-44.
10. G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513-523, 1988.
11. X. Yin, J. Han, and Philip S. Yu. Truth Discovery with Multiple Conflicting Information Providers on the Web. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 20(6):796-808, 2008.
12. X. Jin, S. Spangler, R. Ma, and J. Han. Topic Initiator Detection on the World Wide Web. In *Proceedings of the 19th international conference on World wide web*, pp. 481-490, 2010.
13. N. Kasabov and Q. Song, "DENFIS: Dynamic evolving neural-fuzzy inference system and its application for time-series prediction," IEEE Trans. Fuzzy Syst., vol. 10, no. 2, pp. 144–154, Apr. 2002.
14. P. P. Angelov and X. W. Zhou, "Evolving fuzzy-rule-based classifiers from data streams," IEEE Trans. Fuzzy Syst., vol. 16, no. 6, pp.1462–1475, Dec. 2008.
15. W.Wang and J. J. Vrbanek, "An evolving fuzzy predictor for industrial applications," IEEE Trans. Fuzzy Syst., vol. 16, no. 6, pp. 1439–1449, Dec. 2008.
16. S. Abe and M.-S. Lan, "A method for fuzzy rules extraction directly from numerical data and its application to pattern classification," IEEE Trans. Fuzzy Syst., vol. 3, no. 1, pp. 18–28, Feb. 1995.
17. http://en.wikipedia.org/wiki/Membership_function_(mathematics), 2015.
18. J. Allan. Topic Detection and Tracking: Event-Based Information Organization. Norwell, MA: Kluwer, 2000.
19. Q. Mei and C. Zhai. 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, 2005.
20. C. Wei and Y. Chang. Discovering event evolution patterns from document sequences. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 37(2):273–283, 2007.
21. Y. Jo, C. Lagoze, C. Lee Giles. 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, 2007.
22. R. Nallapati, A. Feng, F. Peng, and J. Allan. Event threading within news topics. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pp. 446-453, 2004.
23. Altman, Alon; Moshe Tennenholtz (2005). "Ranking Systems: The PageRank Axioms" (PDF). Proceedings of the 6th ACM conference on Electronic commerce (EC-05). Vancouver, BC. Retrieved 29 September 2014.
24. J. Himberg, K. Korpiaho, H. Mannila, J. Tikanmaki, and T. Toivonen. Time series segmentation for context recognition in mobile devices. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, 2001.