Journal of Web Engineering, Vol. 14, No.5&6 (2015) 525-544
© Rinton Press

BAYESIAN-BASED TYPE DISCRIMINATION OF WEB EVENTS

QICHEN MA XIANGFENG LUO JUNYU XUAN HUIMIN LIU

School of Computer Engineering and Science, Shanghai University, Shanghai, China

{qichenma, luoxf, xuanjunyu and hliu}@shu.edu.cn

Received February 15, 2015
Revised June 23, 2015

There are a large number of web events emerging on the web and attracting people's attention every day, and it is of great interest and significance to distinguish the different types of these web events in practice. For example, the distinguished emergent web events should be paid more attentions by the departments of the government to save lives and damages or by news websites to increase their hit-rates using limited resources. However, how to efficiently distinguish the types of web events remains a challenge issue due to the seldom efforts paid to this issue in the community. In this paper, we conduct a thorough consideration on this problem and then propose an innovative Bayesian-based model to distinguish the different types of web events. To be specific, all web events are firstly assumed within three types whose formal definitions are given by considering their properties. Aiming to sufficiently describe and distinguish three types web events, a set of specially designed features are then extracted from the volume and the content of web events. Finally, a Bayesian-based model is proposed based on the designed features. The experimental results demonstrate the capability of the proposed model to distinguish types of web events, and the comparisons with other state-of-the-art classifiers also show the efficiency of the proposed model.

Key words: topic detection and tracking, event classification, Bayesian model, web mining

Communicated by: M. Gaedke & O. Diaz

1 Introduction

With the rapid development and broad prevalent of the web, it is almost that every kinds of information in the society can be found on the web. One important content of the web is the web events which are reflections of human activities in the society, such as '*Japan Nuclear Leakage*', '*World News Phone Hacking Scandal*', and '*Volcanic Eruption in Iceland*'. After the emerging of each web event, there will be plenty of webpages published by the journalists or ordinary peoples to report or discuss this event on the web. So the investigation of these web events could help us better understand human behaviours and provide different kinds of services based on that. Although there are a large number of web events every day, they normally do not receive same treatments from people due to different properties of events. Some of them are just routines of daily life and will not attract much attention for a long term; on the contrary, some of them contain emergent or interesting information relevant to the people which have the capability to attract broad attentions in a long period. For example, '*World News Phone Hacking Scandal*' is a breaking news for the British people, so there are plenty of webpages published to report or discuss it after its emerging. Contrarily, there will be little webpages to report '*A Car Accident*' because it is just a normal thing and does not contain any interesting information for the public. Furthermore, the number of web events all over the world could be hundreds of thousands in each day, and the number of webpages of these web events may be millions. Therefore, the automatic way is a must for the analysis of these web events. One significant task of the web event analysis is to automatically distinguish their different types from each other.

The ability to automatically distinguish the types of web events can be used in a variety of settings. For example, 1) among all the events, some emerging web events may give rise to the riots if the government has any delay on the reactions. So the automatic type distinguish from large-scale web events could help the departments of the government pay their limited time and energy to the emergent web events only; 2) the news websites have been trying to attract web users' attentions by ranking news events appropriately on their limited frontpages. So the automatic type distinguish of web events could help news websites to design the positing strategy; 3) at the same time, the news websites should normally maintain a limited number of web servers to crawl the webpages of web events all over the world. Due to the limitation of the crawling ability, the crawled web events should be carefully selected and it is apparently that the automatic type distinguish of web events is helpful for the crawling events selection.

As opposite to the great significance of the task of automatic type distinguish for web events, there is little attention paid to this task in the research community. The most related area for this task is Topic Detection and Tracking (TDT) [1-4] which involves the unknown event detection, information gathering and segmentation, time detection for the event happened, and the detection of following-up report of events [5-8]. Different from our web events type discrimination method, it mainly aims to discover and track the events from the large-scale webpages [31-34]. It remains people's responsibility to distinguish the different types of these discovered or tracked events. Besides, many researchers are studying on the classification and clustering of web text [34] and web knowledge [37] related to web events [35-36], while seldom studies have been conducted on the type discrimination of web events. Some text classification/clustering algorithms [27-29], which are currently considered as the efficient tools for the web event analysis, might be adopted to resolve the web events type distinguish task. However, these algorithms are only based on keyword features to evaluate the semantic similarity and there is no other features considered to distinguish the types of web events, such as temporal features. With the keywords as the only features, these algorithms cannot achieve good performances for this task.

In this paper, we firstly give a thorough consideration on the task of discrimination of web events and formalize the whole problem, then propose a Bayesian-based model to automatically distinguish the web event types from each other. According to their social activity natures, we firstly categorize all the web events into three types, including emergency event, popular event and general event. Each event type gives a formal description to show its characteristics. In order to distinguish three types, we specially design a set of features to capture and distinguish the natures of these types' web events. These features are different from the text clustering/classification algorithms (features are just keywords). Some temporal features which are capable to capture the evolving potentials of the web events are also designed. Even for the traditional keyword features, we have designed different evaluations by the volume change and distribution change. Finally, a Bayesian-based model is proposed based on the designed features.

The main contributions are summarized as follows:

1. We have categorized all web events into three types and given them formal definitions and nature descriptions;
2. We have innovatively designed a set of features to distinguish different types of web events, and the thorough statistical tests have been conducted to show their relative effectiveness;
3. We have proposed a Bayesian-based model to automatically distinguish the types of web events based on the designed features and it achieves good performance on this task.

The rest paper is organized as follows. In section 2, we give the definition of the types of web events. In section 3, we introduce the designed features that may impact on the type discrimination of web event. Based on these features, we propose the Bayesian based function for the type discrimination of web events in Section 4. In section 5, we discuss the independent of features. In section 6, we give the results of experiment on the real-world dataset. In last section, we give a conclusion of our work.

2 Type Definition of Web Events

The different types of web events are formally defined in this section, and the discussion of their features is also introduced.

2.1 The Relation between Web Events and Social Events

The web events come from two sources. One is social events information which can be imaged as the webpages on the web. By the imaging, social events information can spread and evolve on the web, and the web can also feedback the changed information to the society. Interactive feedback of the event information between the society and web leads to the evolution of social events. Such events we call web social events. For example, in July 4, 2011, ‘*News of the World*’ was revealed that illegally wiretapped the phone of the missing girl Millie Dowler and her families in 2002 which led to police involved. This event caused a great repercussion in British, and then a succession of eavesdropping scandals was reported via the social media on the web. The results shocked the world. That scandal spread in the web and was reported by different social media and made citizens worrying about their privacy, and all these led to the scandal breakout. This event reflects that the evolution of social event is deeply influenced by the social media on the web [9].

Another source is the public sentiment on the web. This kind of event does not happen in real world, but do have the influence on the society by the web and may form a social event eventually. Such event we call it public sentiment event. For example, in India in July 2012, a series of rumours and threats spread by SMS, web and other media, causing panic in the entire region. Messages showed local Muslim would begin a large-scale massacre of retaliation, and some web social media and mobile phones spread the pictures show a number of tragic stories of victims. Finally, more than 300,000 people left for a safe place. This typical public sentiment event on the web formed an emergent social event eventually. This reflects that the event occurred on the web also has the ability to influence the society [10]. In this paper, we focus on the events happened on the web or occurred in the society but image on the web.

2.2 The Type Definition of Web Events

Definition 1. The Type of Web Event, ε**:** The type hypothesis space of web event is the collection of all events that may exist on the web, $\varepsilon = \{\varepsilon_1, \varepsilon_2, ..., \varepsilon_n\}$, $\varepsilon_i \subseteq E$; where $E = \{e_1, e_2, ..., e_s\}$ is the collection of web event e; ε_i is one type of web event; n is the number of types.

In this paper, according to the emergency degree, we classify web events into three types. $\varepsilon = \{\varepsilon_1, \varepsilon_2, \varepsilon_3\}$, where ε_1 is emergency events, ε_2 is popular event, and ε_3 is general event.

Definition 2. Emergency event, ε_1**:** Emergency event is the event caused by major natural disasters, accidents, or social security imaged on the web that required to be paid more attention by government or social groups within a special time interval. It also can be a public sentiment event on the web which has a great impact on the society.

Therefore, quick response to the emergency event is very important, or it will cause great negative impact on the society and even lose control. Such as "*5.12 Wenchuan Earthquake*", "*9.11 terrorist attacks*" and "*the British eavesdropping event*". Such an event usually has features of sudden, complexity, destructive, persistent.

Definition 3. Popular event, ε_2**:** Popular event is the social event imaging on the web or public sentiment event occurred on the web. These events are related to people's daily life and people concern it for a long term.

This type of events has been the focus of public, for example, "*Price regulation of house*", "*Food security*" and "*Huangyan Island incident*".

Definition 4. General Event, ε_3**:** General event is the social event imaged on the web or public sentiment event occurred on the web that gets less attention and last a very short time. Such an event usually reported after occurs, and then forgotten by people quickly.

For example, "*Super moon*" and "*Forbes Chinese rich list*" are the typical general events.

3 Features for Type Discrimination

3.1 Features used in Type Discrimination

In order to study the time series of web events, we extract some statistically related features. These features describe the different aspects of web event and represent different significant. A web event has a complete break power in its life cycle L_e , and the time series of break power is represented as: $S = \{s_1, s_2, ..., s_n\}$, where s_i is the break power at time t_i , n is the length of time series. By analyzing the data of time series of web events, we propose six temporal features that useful for the type discrimination of web events.

Outbreak power of web event: Outbreak power is the most basic feature that describes the evolution course of web events. In addition, we can get values of all other features by calculation of the outbreak power of web event. If the outbreak power is higher, the event is more likely to be an emergency event; if the outbreak power is lower, the event is more likely to be a general event.

Average outbreak power of web event: Average outbreak power describes a general level of an event outbreak power in time interval $[t_i, t_j]$. It is also an important basic temporal feature to calculate other temporal features. If a web event has higher average outbreak power, then it is more likely to be an emergency event than to be a general event.

Fluctuation power of web event: For every event, its urgent degree may be changed by interaction between web events and social events in their evolution course, which means its outbreak power changes in different time interval. So the amplitude of the curve of outbreak power also changes in different time interval, and we use fluctuation power to describe these changes. A higher fluctuation the event has, the higher probability of the event to be an emergency event.

Coefficient of skewness of web event: In general, an emergency event has three basic states in its life cycle, i.e., latency state, outbreak state and decline state. In different state its outbreak power curve has different shapes. In latency state, outbreak power curve is flat; in outbreak state, outbreak power curve changes a lot and in decline state, the curve tends to flat. So the curve distribution of outbreak power should have one or more skewnesses instead of a symmetrical distribution. For an event, if the skewness of the distribution does not exist, it may likely to be an emergency event.

Coefficient of kurtosis of web event: Coefficient of kurtosis describes the degree of kurtosis of outbreak power. Variance has directly relationship with coefficient of kurtosis, which means large change of outbreak power leads to high coefficient of kurtosis and generates some peak points. So if the outbreak power distribution changes a lot in different time intervals and has a sharp distribution, then this event is more likely to be an emergency event. But if the distribution of outbreak power is a flat peak distribution, the event is more likely to be an emergency event.

Outliers of web event: Emergency event generally have the possibility of deriving other events. When the derivation happened, emergency event usually has a higher outbreak power in that time. We call it outlier. So in its evolution course, emergency event often has some outliers. If there is no outlier in evolution course of an event, that event is more likely to be a popular event or a general event. On the contrary, if the event has many outliers in its evolution course, then this event may be an emergency event. So we should measure the outliers of web event.

Since temporal features 2)-6) are based on temporal feature 1), feature 1) is redundant and has a strong dependence with other five features. So feature 1) don’t needs to participate in operation, and plays a role by features 2) -6). Whether redundancy exits between features 2)-6) and whether there is a strong interdependency among features 2)-6), we will discuss these questions in the later sections.

3.2 The Definition of Features of Web Event

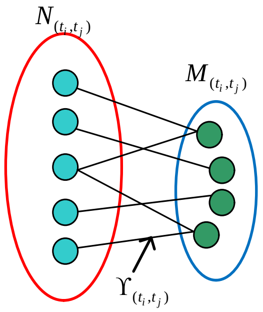


Fig. 1 The relation of three features (i.e., increased web pages $N_{(t_i,t_j)}$, event attributes $M_{(t_i,t_j)}$ and Event attribute distribution in the increased web pages $\Upsilon_{(t_i,t_j)}$) in calculation of outbreak power

According to the above discussion, we need to measure and define these six features. First of all, before building the function of the type discriminant of web event, we need a temporal feature to describe its evolution course, namely outbreak power of web event.

Definition 5. Outbreak power of web event, $\mathrm{op}(t_i,t_j)$ Outbreak power is proportional to the number of increased web pages, the number of increased attributes, and distribution of increased attribute in the increased web pages. It can be denoted as,

$$op(t_i,t_j)\infty\{N_{(t_i,t_j)},M_{(t_i,t_j)},\Upsilon_{(t_i,t_j)}\}$$

A time interval with higher outbreak degrees will have a greater probability as the milestones and peak of web event. Three features involved in the definition 5 are described as follows:

Number of increased web pages $N_{(t_i,t_j)}$: We know if a web event has a vast number of increased web pages in time interval $[t_i,t_j]$, its outbreak power is high. Then it is more likely to be an emergency event.

Number of event attributes $M_{(t_i,t_j)}$: In the case of a certain number of increased web pages in a time interval, the more the event attributes, the higher the breakout power. Then the content of the event involves more and it is more likely to be an emergency event.

Attribute distribution of web event in the increased web pages $\Upsilon_{(t_i,t_j)}$: When the attributes are all in the increased web pages in time interval $[t_i,t_j]$, then these web pages are not the innovative. In this case, people only take care of the event rather than actively discuss it. So its outbreak power is low and it is less likely to be an emergency event. The relation of the three features is showed in Fig 1.

We can use algorithm proposed in [35] with Fig. 1 to calculate outbreak power. The result of calculation describes the evolution course of web events, for example, Fig 2 shows the outbreak power of event "*Japan nuclear leakage*".

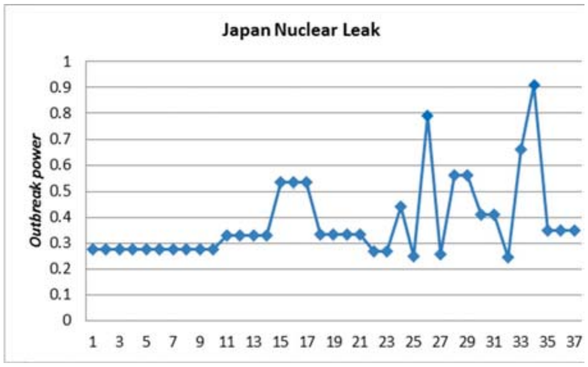


Fig. 2. Timing data observations of event "*Japan nuclear*

After measuring the outbreak power of web event in time interval $[t_i, t_j]$, we need to calculate its average outbreak power. If the average outbreak power is high, the event is more likely to be an emergency event and less likely to be a general event.

Definition 6. Average outbreak power, op_{ave}

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

where $s_i = op_{t_i}$ represents outbreak power of an event at time t_i ; n is the length of the life course of the event.

Average outbreak power op_{ave} is the description of urgent degree in a certain time interval. Web events with high outbreak power may be not emergency events; they can also be popular events. So we need calculate fluctuation power, web event with high fluctuation power is more likely to be an emergency event.

Definition 7. Fluctuation Power, fp

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

where $s_i = op_{t_i}$ represents the outbreak power of event at time t_i ; op_{ave} is the average outbreak power, n is the length of life course of the event.

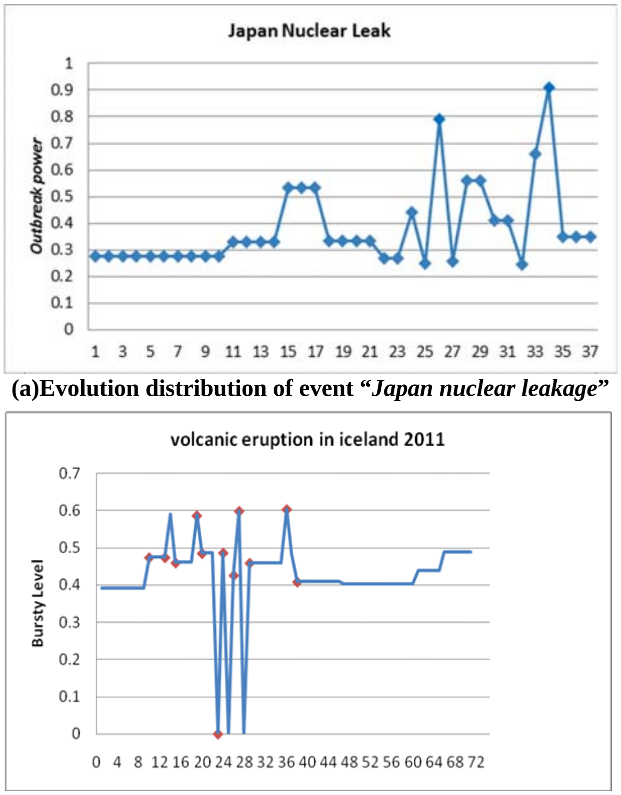
Fluctuation power fp reflects the fluctuation of outbreak power of event in a certain time interval. In general, the outbreak of an emergency event contains three basic states: latency state, outbreak state and decline state. So the curve of outbreak power has much skewness. For measuring skewness, we should define coefficient of skewness.

Definition 8. Coefficient of Skewness, S_k

$$S_k = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{s_i - op_{ave}}{Var})^3 \;,\; Var = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(s_i - op_{ave})^2}$$

where $s_i = op_{t_i}$ represents the outbreak power of event at time t_i; op_{ave} is the average outbreak power; n is the length of life course of the event.

Coefficient of skewness S_k reflects skewness distribution of a web event in a certain time interval. If $S_k = 0$, the distribution is symmetric. If $S_k < 0$, the distribution is left-skewed and has an elongated right tail. Then the event breaks out in the early period of an event, and the event has shorter latency state. As shown in Fig 3(b). If $S_k > 0$, the distribution is right-skewed and has an elongated left tail. Then the event breaks out in the late period, and the event has longer latency. As shown in Fig 3(a).



(a)Evolution distribution of event “*Japan nuclear leakage*”

(b)Evolution distribution of event “*the volcanic eruption in Iceland*”

Fig. 3. Evolution state distribution coefficient of skewness of two web events

Definition 9. Coefficient of Kurtosis, K_f

$$K_f = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{s_i - op_{ave}}{Var})^4 \ , \ Var = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(s_i - op_{ave})^2}$$

where $s_i = op_{t_i}$ represents the outbreak power of an event at time t_i; op_{ave} is the average outbreak power; n is the length of life course of event.

Kurtosis K_f reflects the distribution is peaky or flat of a web event in a certain time interval. $K_f = 3$ is the standard normal distribution. If $K_f > 3$, the distribution shows a peak distribution, as shown in Fig 4(a). Web event has one or more significant peaks in a certain time interval and is more likely to derive sub-events. If $K_f < 3$, the distribution shows a flat distribution, as shown in Fig 4(b), web event does not have obvious peaks in a certain time interval and is less likely to derive sub-events.

If web event has more outliers, it is more likely to be an emergency event. If web event has no outliers, it is more likely to be a popular event or genera event. So we should measure whether there are any outliers in the evolution course of the web event. We give the following definition:

Definition 10. Outliers

$$T_i = \frac{|\, s_i - med(S)\,|}{med\,|\, s_i - med(S)\,|}$$

where T_i represents the fraction T of attribute i; $med(S)$ is the median of time series data; $s_i = op_{t_i}$ represents the outbreak power of event at time t_i.

Any T>5 points called outliers; we select 5 as the threshold because of the probability of this value is approximately 0.001. Outlier implies anomalies happen in the evolution course of web event, that point is the turning point or mutations point.

4 Baysian based Type Discrimination Function of Web Events

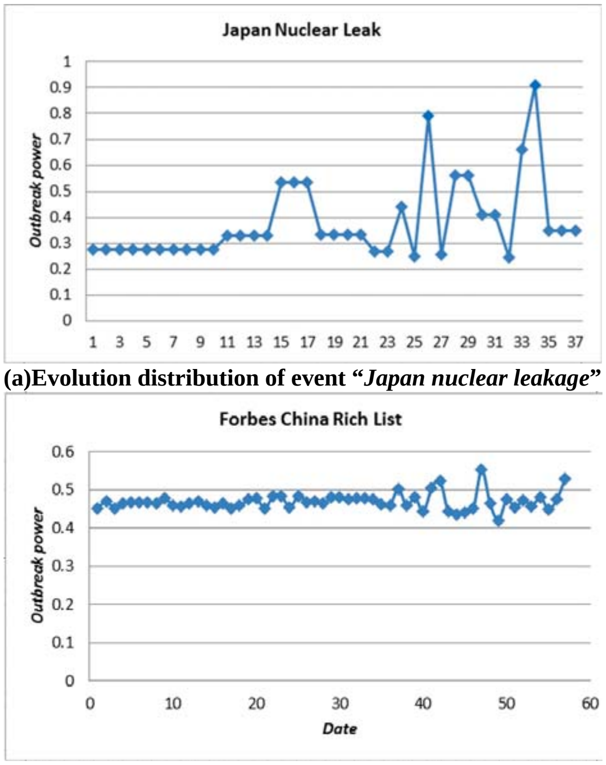
Bayesian model, which is a kind of probability based classifier, is used for subject classification. In this paper, the key idea [14-16] is to classify events into three types in the training set and calculate the priori pattern of each type, and then have the type discrimination with the test data.. Bayesian based the function of type discriminant of web event has the following features:

(1) Bayesian based type discrimination function does not classify an event as a certain type, but recalculate the probability of belonging to each type. The type with maximum probability is the event should belong to.

(2) In general, not just one or few features play a role in type discrimination, but all features play a role.

(3) Features of type discrimination can be discrete, continuous and mixed.

In the previous section, we have introduced some features of web event, such as average outbreak power op_{ave}, fluctuation power fp, coefficient of skewness S_k, coefficient of kurtosis K_f and outliers T. For each web event, their time series data can be expressed as a set of features $T_d(e)=\{op_{ave}, fp, S_k, K_f, T\}=\{D_1, D_2, D_3, D_4, D_5\}$. The value of each feature D_i can be calculated directly from the time series data and the real time data. Different types of events should have different feature patterns. Therefore, we establish the image from type hypothesis space of event to observed values of data, i.e. conditional probability model of event type discrimination, it can be represented as,



(a)Evolution distribution of event "*Japan nuclear leakage*"

(b)Evolution distribution of event "*Forbes Chinese Rich List*"

Fig. 4. Distribution coefficient of Kurtosis of web event

$$P(T_d(e)\,|\,\varepsilon_i)=\prod_{j=1}^{5}P(D_j\,|\,\varepsilon_i) \qquad (1)$$

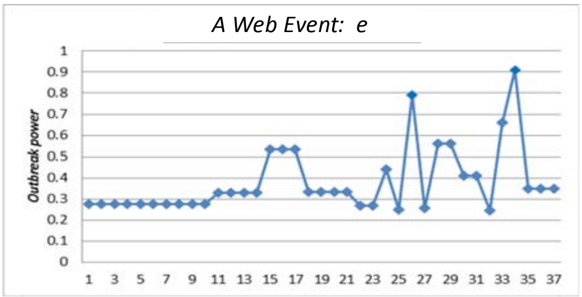
where $P(T_d(e)\,|\,\varepsilon_i)$ is conditional probability model of event type discrimination. Eq. 1 is the model of time series data $T_d(e)$ on the condition of assuming event e belongs to type ε_i. For each feature D_j, $P(D_j\,|\,\varepsilon_i)$ represents the probability of feature D_j with a certain value on the condition of feature D_j belongs to type ε_i, and $P(D_j\,|\,\varepsilon_i)$ called conditional probability.

By the statistics, we can calculate the probability of the event belongs to different types, and establish a probability model of event type discrimination. So we can discriminate the type of unknown event. Classic Bayesian model formula:

$$P(\varepsilon \mid d) = \frac{P(d \mid \varepsilon) * P(\varepsilon)}{\sum_{\varepsilon_i \in \varepsilon} P(d \mid \varepsilon_i) * P(\varepsilon_i)} \qquad (2)$$

For an unknown type of web event, if we get its time series data $T_d(e)$ and according to Eq.2, we can calculate the probability of event e belongs to different type ε_i, denoted as $P(\varepsilon_i \mid T_d(e))$, which is the posterior probability of Bayesian model.

The type ε_i has the largest value of posterior probability is the type that the event belongs to, namely $(\exists \varepsilon_i \to P(\varepsilon_i \mid T_d(e))_{\max}) \to (e \in \varepsilon_i)$. Fig .5 shows the type discrimination of web event.



$$T_d(e) = \{op_{ave}, fp, S_k, K_f, T\} = \{D_1, D_2, D_3, D_4, D_5\}$$

Feature	D_1	D_2	D_3	D_4	D_5
Value	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5

Fig .5. Features involve in type discrimination of web event

Embedding the defined temporal features into Eq.2, we can calculate the probability $P(\varepsilon_i \mid T_d(e))$ of type ε_i that event belongs to:

$$\begin{gathered} P(\varepsilon_i \mid T_d(e)) = P(\varepsilon_i) * \prod_{j=1}^{5} P(D_j = \phi_j \mid \varepsilon_i) / \sum \\ \sum = \sum_i P(\varepsilon_i) * \prod_{j=1}^{5} P(D_j = \phi_j \mid \varepsilon_i) \end{gathered} \qquad (3)$$

According to the discrimination principle of choosing maximum probability, the type ε_i has the largest value is the type that event belongs to.

5 Independence of Temporal Features

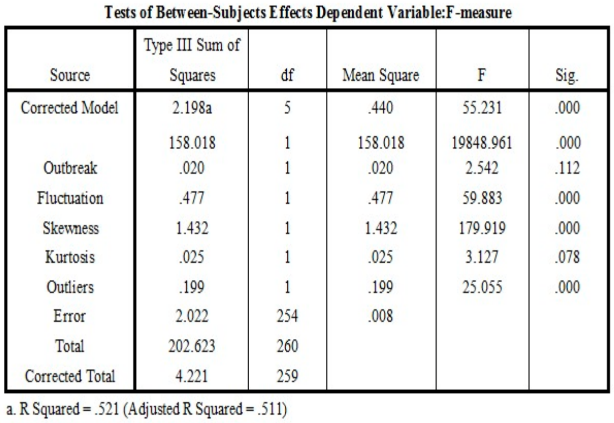
After the establishment of probabilistic model, we need to test the inter-dependence between discriminant function and its features. We must ensure the proposed temporal features are independent with each other in order to use the type discrimination effectively. We also need to verify the validity of the features and remove the features have strong dependence and noise.

5.1 Dependent Test of Temporal Features

If inter-dependence exits among features, it has great impact on discrimination function. Therefore, we expect to remove the features which have strong dependence.

The main idea of factor test is from [18], namely Multivariate analysis. Multivariate analysis is used to study whether two or more variables have a significant impact on the observed variables. We use multivariate analysis of variance [19] (or "F-test") to study a number of factors' influence on observed variables. Multivariate analysis of variance is not only able to analyze a number of factors independent influences on the observed variables, but also analyze the inter-influence on the distribution of observed variables. And ultimately find optimal combination for the observed variable.

In this paper, various features are treated as different factors that affect result of type discrimination. These factors (temporal features) play a role in type discrimination. By multivariate analysis method, we can find the influence of each feature on the model and inter-influence of various features. We can find the optimal combination of features for the type discrimination, thereby optimize our probability model.



Tests of Between-Subjects Effects Dependent Variable:F-measure

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2.198a	5	.440	55.231	.000
	158.018	1	158.018	19848.961	.000
Outbreak	.020	1	.020	2.542	.112
Fluctuation	.477	1	.477	59.883	.000
Skewness	1.432	1	1.432	179.919	.000
Kurtosis	.025	1	.025	3.127	.078
Outliers	.199	1	.199	25.055	.000
Error	2.022	254	.008		
Total	202.623	260			
Corrected Total	4.221	259			

a. R Squared = .521 (Adjusted R Squared = .511)

Fig .6. Each parameter' influence on type discrimination model

5.2 Model Test of Type Discrimination

Experimental design

We select several web events and classify them into three types by manual annotation. These classified events are considered as our experimental data set. From each type of event, we randomly aliquot some events (about 2/3 of all), which are put together as training set. The remaining portion (about 1/3 of all) constitutes the test set.

In the experiment, we use training set and probability model; combine with statistical methods to calculate the conditional probability of each type. Next we use trained probability model to test the events in test set one by one, and then we use accuracy, recall and F-measure to evaluate the results of type discrimination. For two ways of each feature (i.e., add feature or remove feature), we perform the above-mentioned process (constructing data set, training model, testing results and evaluation). Given to random errors, we repeat the process ten times.

Tab 1. The Classification results of different combinations

Group 1	**Group 2**	**Repeat count**
0.973	0.973	1
0.882	0.882	2
0.944	0.914	3
1	1	4
0.914	0.849	5
0.973	0.944	6
1	0.849	7
0.973	0.944	8
0.973	0.849	9
0.914	0.882	10

Experimental results and analysis

By multivariate analysis of variance (F-test) of SPSS (Statistical Product and Service Solutions), we test influence of each feature on type discrimination function, inter-influence between these features' interaction and synergy. As shown in Fig .6, average outbreak power and coefficient of kurtosis these two temporal features do not significantly affect the type discrimination (Sig> 0.05); coefficient of skewness, outliers and fluctuation power, each of them does have significant impact on the type discrimination (Sig <0.05).

Tab2. t-test: Paired two-sample mean analysis

	Group 1	**Group 2**
average	0.955	0.909
variance	0.005	0.003
observed value	10	10
Poisson correlation coefficient	0.415	
mean difference	0	
Df	9	
t Stat	2.761	
P(T<=t) One-tailed	0.011	
t Critical one-tail	1.833	
P(T<=t) two-tailed	0.022	
t Critical two-tail	2.262	

For different combinations of features, we also test their inter-influence on type discrimination model. Then we can get the optimal temporal features combination<para1, para2…>. With the optimal features combination, the noise and redundancy can be removed. Subsequently, we conduct a multivariable analysis of features. Result show: second-order combination case, besides < average, outliers >, < fluctuation, outliers >, < kurtosis, outliers >, < skewness, outliers >, other combinations all have influence on the type discrimination; for third-order combination case, only combination < average, fluctuation, kurtosis > have significant influence on the type discrimination; for high-order combination case, no combination has influence on the discrimination. So the combination < average, fluctuation, kurtosis > is the optimal combination and they are independent with each other.

For testing the results of multivariable analysis, we use combination < average, fluctuation, kurtosis > (labeled as group 1) to do the same experiment again. Then compare with result of combination < average, fluctuation, kurtosis, skewness, outliers > (labeled as group 2). We use paired data t to test our assumption. Table 1 shows the results of two combinations. Repeat the test ten times to reduce errors.

From the table 1 and table 2, we see that the combination< average, fluctuation, kurtosis >has higher accuracy. p (One-tailed) <0.05 means that the combination significantly improves the accuracy of type discrimination. Therefore, the test verifies that redundant features do have dependencies with chosen features and made noise.

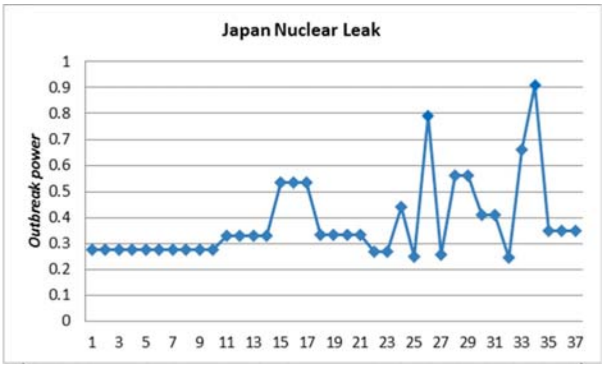


Fig .7. Timing data observations of event "*Japan nuclear leakage*"

6 Discussion on Experiments

In the experiments, we selected a number of typical web events and used the discrimination function to classify them. In addition, we selected three optimized features to construct the function of type discrimination of web events, and compared with the function constructed by other redundant features. It proved that the optimized features can reduce computational complexity, and also can improve the accuracy of type discrimination.

6.1 Web Event Instances

The first case study is the event of "Japan nuclear leakage"

The time data observation was shown in Fig .7.

According to the Eq.3, using all temporal features, posterior probabilities of each type can be calculated.

$$P(\varepsilon_1 | D) = P(\varepsilon_1) * P(D | \varepsilon_1) / \sum \ =0.995$$

$P(\varepsilon_2 \mid D) = P(\varepsilon_2) * P(D \mid \varepsilon_2) / \sum$ =0.002

$P(\varepsilon_3 \mid D) = P(\varepsilon_3) * P(D \mid \varepsilon_3) / \sum$ =0.003

So it can be inferred that event "*Japan nuclear leakage*" should belong to emergency event (ε_1).

By using feature combination <average, fluctuation, kurtosis >, we calculated posterior probabilities of each type.

$P(\varepsilon_1 \mid D)$ =0.977

$P(\varepsilon_2 \mid D)$ =0.002

$P(\varepsilon_3 \mid D)$ =0.021

So event "*Japan nuclear leakage*" belongs to emergency event (ε_1).

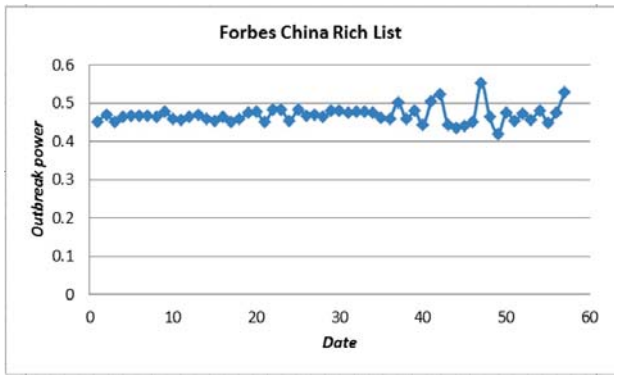


Fig .8. Timing data observations of event "*Forbes Chinese Rich List*"

The second case study is the event of "Forbes Chinese Rich List"

The time data observation was shown in Fig .8.

According to Eq.3 and using all features, posterior probabilities of each type can be calculated as follows.

$P(\varepsilon_1 \mid D) = P(\varepsilon_1) * P(D \mid \varepsilon_1) / \sum = 2.42E\text{-}5$

$P(\varepsilon_2 \mid D) = P(\varepsilon_2) * P(D \mid \varepsilon_2) / \sum = 0.008$

$P(\varepsilon_3 \mid D) = P(\varepsilon_3) * P(D \mid \varepsilon_3) / \sum = 0.992$

Based on the above probabilities, we know that the event "*Forbes Chinese Rich List*" belongs to general event (ε_3).

By using feature combination <average, fluctuation, kurtosis >, we also can calculate the posterior probabilities of each type.

$P(\varepsilon_1 \mid D) = 3.741E-4$

$P(\varepsilon_2 \mid D) = 0.091$

$P(\varepsilon_3 \mid D) = 0.909$

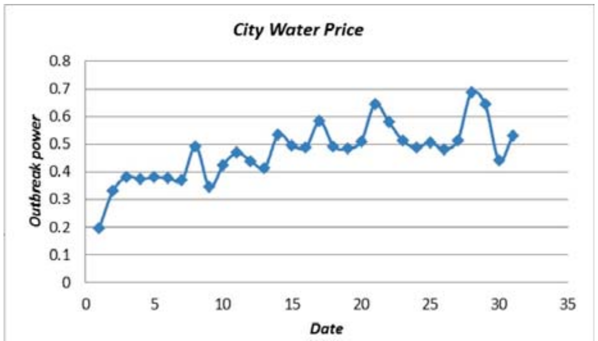


Fig .9. Timing data observations of event "*Price of City Water*"

So the event "*Forbes Chinese Rich List*" belongs to general event (ε_3). According to Fig .8, we also know this event belongs to emergency event (ε_3).

The third case study is the event of "Price of City Water"

The time data observation was shown in Fig .9.

According to Eq.3, by using all features, posterior probabilities of each type can be calculated.

$P(\varepsilon_1 \mid D) = P(\varepsilon_1) * P(D \mid \varepsilon_1) / \sum \ = 0.002$

$P(\varepsilon_2 \mid D) = P(\varepsilon_2) * P(D \mid \varepsilon_2) / \sum \ = 0.462$

$P(\varepsilon_3 \mid D) = P(\varepsilon_3) * P(D \mid \varepsilon_3) / \sum \ = 0.536$

So event "*Price of City Water*" belongs to general event (ε_3). By using feature combination <average, fluctuation, kurtosis >, we also can calculate posterior probabilities of each type.

$P(\varepsilon_1 \mid D) = 2.325E - 4$

$P(\varepsilon_2 \mid D) = 0.759$

$P(\varepsilon_3 \mid D) = 0.240$

So the event "*Price of City Water*" belongs to general event (ε_3). Referring artificial mark and definition, this event usually should be classified as popular event. According to Fig .9, we can know this event belongs to popular event (ε_2). From the above examples, we can see using the optimized features combination can reduce the complexity of the algorithm while the accuracy is also got to keep. For each unknown event, we get its features, using type discrimination function to calculate posterior probability of each type event belongs to. Then determining which type an event should belong to.

7.2 Experimental Verification

In our experiment, we select 100 web events from Baidu(http://news.baidu.com)，Google (http://news.google.com) and other news sites as our experimental data set. These events cover topics of political, accident, disaster, terrorist attacks in various fields and involve 900,000 pages. Table 3 shows, the experimental data set.

Table 3. The details of dataset used to type discriminant of web events (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

From these 100 events, we select some events (about 2/3 of all) as training set. The remaining portion (about 1/3 of all) constitutes the test set. Each event in training set is labelled event type by manual work. By statistics of training set, we can calculate the probability of each feature pattern, which is presented by different types of web events. In addition, we also can make use of the temporal features we have known, and then we can get Bayesian based type discriminant function. At this point, we have established temporal features for type discrimination by means of training set.

Next, we will mine the features for type discrimination in test set. For every web event, we first obtain a set of feature $T_d(e)=\{D_1,D_2,D_3,D_4,D_5\}$; then we calculate the posterior probability $P(\varepsilon_i \mid T_d(e))$ by means of the proposed type discriminant function. Finally according to maximum probability principle, we can determine which type that web event should belong to.

A major evaluation for type discrimination result is that whether the proposed type discriminant function has a high accuracy. For each event in test set, we seek the views of a group of label members to test the result is correct or not and to test the effectiveness of the proposed type discriminant function. For example, the event "*Japan nuclear leakage*" is discriminated as popular event, but artificial label is emergency event, and then we think this discrimination is failed. Furthermore, each label member of labelling events finished reviewing independently to ensure the reliability and effectiveness of the experimental results. Before reviewing, we provide label members with an abstract description of each type of web events. Finally, the results of the evaluation of all label members are bundled together and reach a consensus. As shown in Table 4, the discrimination accuracy rate of three types is more than 85%. Above experiment verifies that the proposed type discriminant function is effectiveness.

Table 5 shows the experiment results of three algorithms. The proposed method has a better performance than the method of logistic and SMO.

From the established type discriminant function, we can see the fluctuation power of general event is low. This result may be due to the nature of general event. Evolution distribution of time series of general events is manifested as amplitude and fluctuations around its mean. On the contrast, the distribution of emergency events and popular events usually present one or several gathered centers. So the fluctuation power of general events is lower than other two types.

Table 4 experiment results of type discrimination

Evaluation	**Type of event**			
	Emergency event	Popular event	General event	
Precision	90.7%	85%	86.7%	88.1%
Recall	85%	87.9%	93.3%	88.7%

Table 5 experiment results of different methods

	Method		
	Bayesian	**logistic**	**SMO**
Precision	88.1%	88%	85.3%
Recall	88.7%	85%	80%

From the experimental data set, we find emergency events and popular events have higher outbreak power. Emergency events have a high degree of outbreak power, such as “*Japan nuclear leakage*”. Popular events are the topics people concern in a long term, such as “*price rising of daily use*”. General events are the topics people less concern or concern in a short period, such as “*super moon*”.

8 Conclusions and Future Work

As opposite to the great significance of the task of automatic type distinguish for web events, there is little attention paid to this task in the community. We have investigated this task from the beginning (i.e., event type defining and feature designing) to the end (model constructing and evaluating). All the web events have been categorized into three types: emergence event, popular event and general event with their nature descriptions. Then, a set of specially designed features have been proposed to distinguish the natures of different web event types. Their inter-dependences have been analysed through the statistical tests. We found that the combination of < average, fluctuation, kurtosis> achieves the best performances among all the features. Finally, a Bayesian-based model has been built to do the type discrimination task. Experiments on the real-world datasets show the efficiency of the proposed model and the superior performance comparing other state-of-the-art methods that are possible to be used for type discrimination task.

The type discrimination in this paper did not consider the time factor. As we know, the type of web events will change with time. A web event may be an emergence event at the beginning stage, but it may lose the interests of the public as its evolution and become a general event. Therefore, how to efficiently detect this type transition will be an interesting and challenge issue. Next, we are going to use Concept Drift techniques to resolve this issue.

Acknowledgements

Research work reported in this paper was partly supported by the National Science Foundation of China under grant no.61471232 and the Key Innovation Program of Shanghai Municipal Education Commission under Grant 13ZZ064. Corresponding author: Xiangfeng Luo is with the School of Computer Engineering and Science, Shanghai University, 333 Nanchen Road, Baoshan District, Shanghai, China, 200444. E-mail: luoxf@shu.edu.cn.

References

1. J. Allan, R. Papka, and V. Lavrenko, "On-line New Event Detection and Tracking", Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia,1998, pp. 37-45.
2. Y.M. Yang, J.G. Carbonell, R.F. Brown, T. Pierce, B.T.Archibald and X. Liu, "Learning Approaches for Detecting and Tracking News Events", IEEE Intelligent System, 1999,14(4), pp. 32-43.
3. J. Allan, J.G. Carbonell, G. Doddington, J. Yamron and Y.Yang, "Topic Detection and Tracking Pilot Study: Final Report", Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Virginia, USA, 1998, pp. 194-218.
4. J. Bengel, S. Gauch, E. Mittur and R. Vijayaraghavan. "Chat track: Chat room topic detection using classification". In 2nd Symposium on Intelligence and Security Informatics, Tucson, Arizona, 2004, pp. 266-277.
5. T. Brants, F. Chen and A. Farahat. "A System for New Event Detection". In Proc. of ACM SIGIR'03, 2003, 330-337.
6. H.Liu, "Internet public opinion hotspot detection and analysis based on K-means and SVM algorithm", 2010 International Conference of Information Science and Management Engineering, pp.257-261.
7. Q.Guan,S.Ye,etc. "Research and Design of Internet Public OpinionAnalysis System", 2009 IITA International Conference on Services Science, Management and Engineering, pp.173-177.
8. X.Li, "The Design and Implementation of Internet Public Opinion Monitoring and Analysis System", 2nd International Conference on e-Business and Information System Security, 2010, pp.1-5.
9. http://en.wikipedia.org/wiki/News_of_the_World_phone_tapping_scandal.
10. http://www.financialinfo.co/the-northeast-india-offsite-great-escape.html.
11. Griffiths, T. L., Kemp, C., &Tenenbaum, J. B. (in press). Bayesian models of cognition. In R. Sun (Ed.),Cambridge handbook of computational cognitive modeling. Cambridge: Cambridge University Press.
12. C. Kemp et al. Learning causal schemata. S. McNamara, J.G. Trafton (Eds.), Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society, Cognitive Science Society (2007), pp. 389–394.
13. Thomas L. Griffiths, Nick Chater, Charles Kemp, Amy Perfors, Joshua B.Tenenbaum. Probabilistic models of cognition: exploring representations and inductive biases. Trends in Cognitive Sciences, Volume 14, Issue 8, August 2010, Pages 357-364.
14. http://en.wikipedia.org/wiki/Bayesian_model.
15. L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In Proc. CVPR, 2005.

16. S. Aksoy, K. Koperski, C. Tusk, G. Marchisio, J.C. Tilton. Learning Bayesian classifiers for scene classification with a visual grammar. IEEE Transactions on Geoscience and Remote Sensing, 43 (3) (2005), pp. 581–589.
17. S. Paek, S.-F. Chang, A knowledge engineering approach for image classification based on probabilistic reasoning systems, in: IEEE International Conference on Multimedia and Expo, vol. II, New York, 2000, pp. 1133–1136.
18. http://en.wikipedia.org/wiki/Multivariable_analysis.
19. http://en.wikipedia.org/wiki/MANOVA.
20. R. Schwartz, T. Imai, L. Nguyen, and J. Makhoul. “AMaximum Likelihood Model for Topic Classification of Broadcast News.” Euro speech ’97, Rhodes, Greece. September, 1997.
21. J. Allan.Topic Detection and Tracking: Event-Based Information Organization. Norwell, MA: Kluwer, 2000.
22. Korb, K., & Nicholson, A. (2003). Bayesian artificial intelligence. Boca Raton, FL: Chapman and Hall/CRC.
23. Ge, X. & Smyth P. (2001). Segmental Semi-Markov Models for Endpoint Detection in Plasma Etching. To appear in IEEE Transactions on Semiconductor Engineering.
24. Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. Cognitive Science, 30, 555–580.
25. Neal, R. M. (1993). Probabilistic inference using Markov chain Monte Carlo methods (Tech. Rep.No. CRG-TR-93-1). University of Toronto.
26. Sloman, S. (2005). Causal models: How people think about the world and its alternatives. Oxford: Oxford University Press.
27. Xiaochun He, Conghui Zhu , Tiejun Zhao . Research on short text classification for web forum. Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on, Page(s): 1052 – 1056.
28. Yulei Zhang , Yan Dang , Hsinchun Chen . Gender Classification for Web Forums. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 41(4) (2011), pp. 668-677.
29. Ayyasamy, R.K.Alhashmi, S.M. ; Siew Eu-Gene .Concept based modeling approach for blog classification using fuzzy similarity. Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on, Page(s): 1007 – 1011.
30. Lee, K.; Palsetia, D.; Narayanan, R.; Patwary, M.M.A.; Agrawal, A.; Choudhary, A.Twitter Trending Topic Classification. Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on , (2011), Page(s): 251 – 258.

31 On B W, Omar M, Choi G S, et al. Gathering web pages of entities with high precision[J]. Journal of Web Engineering, 2014, 13(5-6): 378-404.

32 Keramati A, Jafari-Marandi R. Webpage clustering: taking the zero step—a case study of an Iranian website[J]. Journal of Web Engineering, 2014, 13(3-4): 333-360.

33 Luo X, Xuan J, Liu H. Web event state prediction model: combining prior knowledge with real time data[J]. Journal of Web Engineering, 2014, 13(5-6): 483-506.

34 Han X, Sun L, Zhao J. Collective entity linking in web text: a graph-based method[C]//Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011: 765-774.

35 Xu Z, Chen H Y. Semantic Outbreak Power Based Evolution of Web Event in Large-Scale Ubiquitous Contexts[J]. International Journal of Distributed Sensor Networks, 2014.

36 Wang X, Luo X, Liu H. Measuring the veracity of web event via uncertainty[J]. Journal of Systems and Software, 2014.

37 Li Q, Lau R W H, Wah B, et al. Guest Editors' Introduction: Emerging Internet Technologies for E-Learning[J]. Internet Computing, IEEE, 2009, 13(4): 11-17.