Research Article

Semantic Outbreak Power Based Evolution of Web Event in Large-Scale Ubiquitous Contexts

Zheng Xu\textsuperscript{1,2} and Hai-Yan Chen\textsuperscript{3}

\textsuperscript{1}Tsinghua University, Beijing, China
\textsuperscript{2}The Third Research Institute of the Ministry of Public Security, Shanghai 201142, China
\textsuperscript{3}East China University of Political Science and Law, Shanghai, China

Correspondence should be addressed to Zheng Xu; zhengxu@shu.edu.cn

Received 30 October 2014; Accepted 5 December 2014

Copyright © 2015 Z. Xu and H.-Y. Chen. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Nowadays, emergencies’ events have a great impact on people’s daily life. Web is acting as an important platform for the diffusion and evolution of social events. With the popularization and development of web technology in the world, web is becoming an important platform to cover, transmit, and release the news. Information holders can use Internet as a medium to broadcast various news in real time. But it is difficult for people to excavate the events development trend from web information. In this paper, we introduce a method to measure the evolution process of the Web event based on the semantic outbreak power. Then, we propose an approach to distinguish the event type based on the fuzzy recognition. Finally, we give some instances validation of semantic outbreak power of web event and the correctness verification of the event type distinguish of Web event.

1. Introduction

With the popularization and development of web technology in the world, web is becoming an important platform to cover, transmit, and release the news. Information holders can use Internet as a medium to broadcast various news in real time. Web users can quickly and comprehensively grasp the dynamic events [1].

Web event refers to the web event which is sustainedly and extensively reported and discussed in short online media (such as BBS, news sites, and blogs) [2, 3]. It is likely to cause some harm to the social reality. The reported and discussed forms are various, such as a web page news commentary, the post and reply in BBS, and the records and message in blog. These reports and discussions have a major impact on the social stability and a large number of Internet users. It often accompanies the occurrence, development, and change of the hot topics and social events. Web event has some typical features: (a) it may have wide and speedy transmission; (b) it may have devastating effects on the society; (c) it is not easy to find and control the breaking point. So how to detect and compute the propagation speed, predict the direction of evolution, and effectively control web event is a major challenge. Therefore it is necessary to analyze the future of the web event and measure its evolution.

With the development of internet of things [4–6], big data [7–10], and cloud computing [11–14], evolution is a basic feature of the web events and important research content in the field of topic detection and tracking (TDT) [3]. The main research contents of the TDT [15, 16] involve topic detection and information collection, segmentation of event information and the first report event time detection, and topic tracking. Generally speaking, the TDT technologies try to detect unknown events and cluster news related to these events.

The TDT traces the development of the events. But it is not to measure the dynamic evolution process of the events and does not consider event semantic characteristics in the evolution process. So the TDT cannot offer us a global and clear understanding of the web event. Thus, in order to recover the insufficiency of the TDT, we have proposed a method to measure the evolution process of the web event based on semantic outbreak power. Then, we put forward an
approach to distinguish the event type based on the fuzzy recognition.

The paper is organized as follows. Section 2 presents briefly the topic detection and topic tracking. Section 3 introduces a method to measure the evolution process of the web event based on semantic outbreak power. Section 4 puts forward an approach to distinguish the event type based on the fuzzy recognition. Section 5 gives some instances of validation of semantic outbreak power of web event and the correctness verification of the event type distinguish of web event.

2. Related Work

The TDT [17–19] was first initiated by the Defense Advanced Research Projects Agency (DARPA) and National Institute of Standards and Technology (NIST). It aims to develop a series of information organization technology based events and help people cope with information overload problem. Then, we mainly introduce the research of the two subtasks: topic detection and topic tracking.

2.1. Topic Detection. Topic detection puts data stream from the newswire and news sources reported into different topics. If it is necessary, topic detection will establish the new topic of technology. Topic detection can be thought of as an event cluster. Most of these clusters are incremental [20].

Most of current research topic detection using the traditional method of natural language processing: the center vector method [17], KNN [21], K-means [22], single pass clustering algorithm [23], and so forth. The processing object of the topic detection is the reporting flow, which is changing with time, rather than a static text set. So most studies use a combination of a variety of clustering strategies to extend and improve these basic clustering algorithms.

More representative research includes the following: Yang et al. [24] used the strategy which combined the condensing type clustering algorithm and the average clustering algorithm; the researchers of CMU [25] mainly use the single pass clustering method with time Windows to detect the topic; literature [26] proposed a hierarchical community discovering algorithm based on event network, which exploits the semantic properties of event nodes and edge-weight information in the network, to discover fine granularity communities that are semantically meaningful.

In addition, establishing the model for the topic and report is another important study in the process of topic detection. Related models include space vector model, probabilistic model, lexical chain model, and graph model. The vector space model is the most commonly used model. There are many kinds of the formula to calculate the similarity based on this model, such as okap, Clarity, Hellinger, Tanimoto, Weightsum, and cosine similarity formula.

2.2. Topic Tracking. Topic tracking is the technology used to identify the rest news associated with the topic from news flow according to the given small amounts of training reports associated with a topic. The essence of this task is a guided learning. Topic tracking uses few positive instances data (one or more samples) and a lot of negative instances (history data) to get a classifier. This classifier is used to distinguish whether the new report is associated with the topic. So the topic tracking can be regarded as a special kind of binary classification problem. Many of the technologies in text categorization are the foundation of topic tracking.

In the topic tracking study, the commonly used methods include k nearest neighbours (KNN) [27, 28], the decision TREE (D-TREE) [29], Rocchio algorithm [30], the SVM algorithm [31], and language model [32]. Most of the research improved the basic algorithm and then applied it to the topic tracking. The following simple introduce several representative researches.

Allan et al. used Rocchio algorithm to implement the topic tracking. The core idea of the algorithm is the structural strategy of the topic model. If the feature is conducive to the topic description, its weight will be strengthened. If the feature is tending to wrongly guide topic description, its weight will be weakened [33]. Leek and Sista used the probability model in their topic tracking and recognition system, mainly based on the simple Bayesian algorithm. The system combines multiple classifier system, organized to present the results of each classifier to the user [32].

In addition, another research about web events or topics includes this paper [34] which presents a novel hot event discovery framework to detect hot events online. It contained three stages: document preprocessing, threshold-resilient document classification, and adaptive splitting document clustering. Deng and Xu [35] propose a method to measure the influence and represent the event evolution graph. Lee et al. [36] proposed an incremental tracking framework for cluster evolution over highly dynamic networks. This paper considered the event evolution tracking task in social streams as an application, where a social stream and an event are modeled as a dynamic postnetwork and a postcluster, respectively.

Generally speaking, the traditional TDT technologies just detect unknown events and put related news reports into different topics based on text clustering and text classification. The TDT also traces the development of the events but does not analyze the event deeply. So it cannot provide us with a global and clear understanding of the web event.

3. The Semantic Outbreak Power

3.1. Semantic Features of the Web Event. Our goal is to measure the evolution process of the web event based on semantic outbreak power. So, first of all, we need to get the semantic features set of the web events in the evolution process (Table 2).

Definition 1 (the semantic features set of web event Fe). The semantic features set of web event Fe consists of three parts including event seed set $S(t, t_f)$, web page set $\phi(t, t_f)$, and events keywords set $K(t, t_f)$.

Events seed set $S(t, t_f)$ contains the core attributes word of events which are usually used as the keywords to search for events. Web page set $\phi(t, t_f)$ is a set of resources on the web. The set is a cluster of related web document collections which
Table 1: The new increased web page number of the events of “Libya unrest” every day.

| Date     | News | Blog | BBS |
|----------|------|------|-----|
| 2011-2-22 | 104  | 377  | 616 |
| 2011-2-23 | 132  | 383  | 602 |
| 2011-2-24 | 100  | 781  | 13300 |
|          |      |      |     |
| 2011-3-29 | 3820 | 256  | 745 |
| 2011-3-30 | 4110 | 72   | 367 |
|          |      |      |     |

Table 2: The semantic features original set of the web event evolution process.

| Topic ID | date | $|\Delta \varphi(t_i, t_j)|$ | $|\Delta K(t_i, t_j)|$ |
|----------|------|-----------------|------------------|
| 33       | 0    | 103             | 55               |
| 33       | 0    | 107             | 67               |
| 33       | 1    | 114             | 111              |
| 33       | 1    | 114             | 115              |
| 33       | 1    | 116             | 111              |
| 33       |      |                 |                  |

Table 3: The details of dataset 1 (50 events).

| Feature | Value |
|---------|-------|
| Average number of seeds per event | 2 |
| Average number of web pages per event | 1763 |
| Average number of event attributes per event | 6118 |
| Average number of days per event | 42 |
| Average number of web pages per day | 59 |
| Average number of event attributes per day | 1011 |

Definition 4 (the distribution of event attributes in the new increased web page: $\psi(t_i, t_j)$). For a web event $e$, all the web pages of $\Delta \varphi(t_i, t_j)$ can be used as the vector of $\Delta K(t_i, t_j)$ to represent the event attributes, namely, $\varphi_n = \{w_{1n}, w_{2n}, ..., w_{mn}\}$. $w_{mn}$ represents the event keyword $k_m$ weights of $n$th pages. $m$ is the number of event attributes. These vectors constitute a matrix. The distribution of event attributes in a web page from time $t_i$ to $t_j$ is

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

(1)

The new increased web page set $\Delta \varphi(t_i, t_j)$, the new increased events keywords set $\Delta K(t_i, t_j)$, and the distribution of event attributes in the new increased web page $\psi(t_i, t_j)$ can be obtained from the source data shown in Table 3, and calculate the binary chart shown in Figure 1. The picture reflects the relationship of web page, event keywords, and the distribution of event attributes in the web page.

3.2. The Semantic Outbreak Power

Definition 5 (the outbreak power of web event: $\text{op}(t_i, t_j)$). From time $t_i$ to $t_j$, the outbreak power of web event can be labelled as $\text{op}(t_i, t_j)$. It represents the fact that the severity power of the event $e$ may cause harm and urgency power to deal with emergency during the time $[t_i, t_j]$.

We understand the web event life process more comprehensively after discussing the semantic features set of web events. We will propose an iterative algorithm combining all of these characteristics of events, measuring the web event life process.
Input: The set of pages $\psi(t_i, t_j)$, the set of keyword $K(t_i, t_j)$, the distribution of keywords among pages $\psi(t_i, t_j)$.

Output: The outbreak power $\text{op}(t_i, t_j)$.

for each $\varphi \in \psi(t_i, t_j)$
  
  $c(\varphi) \leftarrow c_0$;  \text{/* setting initial state*/}
  
  $\bar{T} = \{c_0(\varphi_1), c_0(\varphi_2), \ldots, c_0(\varphi_n)\}$;
  
  Repeat
  
  $\frac{\text{er}(k)}{c(\varphi)} \leftarrow c(\varphi)$
  
  $c'(\varphi) \leftarrow \frac{\text{er}(k)}{c(\varphi)}$
  
  $\bar{T}' = \{c'(\varphi_1), c'(\varphi_2), \ldots, c'(\varphi_n)\}$

Until cosine similarity of $\bar{T}$ and $\bar{T}'$ is greater than $\beta$.

Compute $\text{op}(t_i, t_j)$ by (4)

Algorithm 1: The steps of computing the semantic outbreak power.

**Figure 1:** The relationship of the events seed set $S(t_i, t_j)$, the new increased web page set $\Delta \psi(t_i, t_j)$, the new increased events keywords set $\Delta K(t_i, t_j)$, and the distribution of event attributes in the new increased web page $\psi(t_i, t_j) \psi(t_i, t_j)$.

In the previous section, the semantic features set of web events have been defined and attained. Now, we will use these features to calculate the outbreak power. Before proposing the Calculation method, we introduce two important propositions: the representability of event keywords and the credibility of a web page.

**Proposition 6** (the representability of event keywords: $\text{er}(k)$). It describes the ability of the event keywords to express the event.

**Proposition 7** (the credibility of a web page: $c(\varphi)$). It describes the believable event power of the keywords to express the event.

According to the observation of the actual data set and cognitive knowledge, we give the following few deductions to be the basis of calculating semantic outbreak power algorithm (Algorithm 1).

**Deduction 1.** The more web pages on event discussion, the higher semantic outbreak power during the time $[t_i, t_j]$, namely, $\text{op}(t_i, t_j) \propto |\Delta \psi(t_i, t_j)|$.

**Deduction 2.** The more keywords on event discussion, the higher semantic outbreak power during the time $[t_i, t_j]$, namely, $\text{op}(t_i, t_j) \propto |\Delta K(t_i, t_j)|$.

**Deduction 3.** When the disagreement is larger about the event discussion, the distribution of event keywords in the new increased web page $\psi(t_i, t_j)$ is one-to-one mapping. Then, the semantic outbreak power is higher, namely, $(\forall w_{mn} \in \psi(t_i, t_j) \rightarrow w_{mn} \neq 0) \rightarrow \text{op}(t_i, t_j)$. This means that all the webs are different. At this point, the disagreement is bigger and the event is likely to further deteriorate. In contrast, we can make Deduction 4.

**Deduction 4.** If the web page number and keywords on the event discussion are decay during the time $[t_i, t_j]$, the distribution of event attributes in the new increased web page (binary diagram in Figure 1) tends to be a complete graph. At this time, the power of the outbreak of the event is low.

According to Deduction 4, if the entire keywords exit in every web page, namely, the event keywords set and web page set constitute a complete graph, the similarity power between each page will be 1. This means that all web pages are reproduced from a web page. At this point, the web event is relatively uniform and is less likely to generate further threats.

From the previous discussion, we can know that the greatest influences on the semantic outbreak of web events are the new increased web page set $\Delta \psi(t_i, t_j)$, the new increased keywords set $\Delta K(t_i, t_j)$, and the distribution of event attributes in the new increased web page: $\psi(t_i, t_j)$.

Considering the above discussion, we can use the representability of event keywords $\text{er}(k)$ to infer the credibility of a web page $c(\varphi)$. The reverse is also true. It is shown in Figure 2 which consists of three parts: the new increased web page set $\Delta \psi(t_i, t_j)$, the new increased keywords set $\Delta K(t_i, t_j)$, and the distribution of event attributes in the new increased web page: $\psi(t_i, t_j)$.
According to Figure 2, we get the following two deductions.

**Deduction 5.** For a web page \( \varphi \), if most of the event keywords have the strong representability \( er(k) \), the web page will have high credibility \( c(\varphi) \).

**Deduction 6.** If a keyword \( k \) is provided by a higher \( c(\varphi) \) web page, the keyword \( k \) will have a strong \( er(k) \). \( er(k) \) and \( c(\varphi) \) affect each other. So we can use alternative algorithm to compute the semantic outbreak power considering the new increased web page, the new increased keywords, distribution of keywords in the web page.

According to Deduction 4, for a web page \( \varphi \), we can compute the average of the representability of event keywords to compute the credibility of a web page. Consider

\[
c(\varphi) = \frac{\sum_{k \in Kw(\varphi)} er(k)}{|Kw(\varphi)|},
\]

where \( Kw(\varphi) \) is the event keywords set for a web page \( \varphi \).

According to [38], we use the probability function to calculate the representability of event keywords:

\[
er(k) = 1 - \prod_{\varphi \in Wp(k)} (1 - c(\varphi)),
\]

where \( Wp(k) \) is the web page set event keywords \( k \).

The semantic outbreak power \( op(t_i, t_j) \) during the time \( [t_i, t_j] \) is

\[
 op(t_i, t_j) = \sum_{\varphi=1}^{n} (1 - c(\varphi_q)),
\]

where \( n \) is \( |\Delta\varphi(t_i, t_j)| \) and it represents the new increased web pages during the time \( [t_i, t_j] \).

Therefore, we put forward the detailed steps on semantic outbreak power algorithm as follows:

1. give each page the initial credibility; at the same time, all the initial credibility is a vector of a web page \( \vec{T} \); namely, \( \vec{T} = \{c(\varphi_1), c(\varphi_2), \ldots, c(\varphi_n)\} \);
2. according to formula (4), compute the representability of each event keyword \( er(k) \);
3. according to formula (3), compute the credibility of a web page: \( c(\varphi)' \);
4. all types of the credibility calculated using (3) constitute a vector: \( \vec{T}' \), namely, \( \vec{T}' = \{c(\varphi_1)', c(\varphi_2)', \ldots, c(\varphi_n)\}' \);
5. computing will reach a stable state until cosine similarity of \( \vec{T} \) and \( \vec{T}' \) is greater than \( \beta \); at this time, the representability of event keywords \( er(k) \) and the credibility of a web page \( c(\varphi) \) will no longer be changed; if this condition is not satisfied, make \( \vec{T} = \vec{T}' \); turn to step (2); if this condition is satisfied, end the iteration;
6. according to formula (4), calculate the semantic outbreak power.

4. Experiment

**4.1. The Verification of the Semantics Outbreak Power.** Experiments involve events derived from Baidu news sites. We chose 50 web events about 450000 web pages and 100 events about 900000 web pages as the experimental data set, which includes various areas such as political, accidents, disasters, and terrorist attacks. Tables 3 and 4 show the statistical results of the experimental data set in detail in this paper. The determination of these web start events stamp according to the method in [26], and the end events stamp general settings for the experimental process in the time we crawl events related web pages; the average length of each event sampling
Table 4: The details of dataset 2 (100 events).

| Feature                          | Value  |
|----------------------------------|--------|
| Average number of seeds per event| 2      |
| Average number of web pages per event| 5556   |
| Average number of event attributes per event| 16856  |
| Average number of days per event | 40     |
| Average number of web pages per day| 146    |
| Average number of event attributes per day| 469    |

time is 30–40 days. In addition, to determine the evolution process of event, we also obtain semantic futures of events, including event seed set, event web page set, and the event keywords set. We get the event seeds from Baidu. Baidu provides hot issues; at the same time, it also provides hot search words to help users search events.

For example, about Japan’s catastrophic earthquake and tsunami disaster in 2011, many web users want to search the related web page information. Baidu will provide a set of keywords about events such as “Japan, earthquake, tsunami.” We will be event seed as a search term and crawl events related web from the Internet, after getting the event seed from the search engines.

The detailed steps we collect from the web resources are

1. obtaining event seed set \( S(t_i, t_j) \) of web event from the news websites Baidu, such as “Japan, earthquake, tsunami,”
2. using the event seed set as the search keywords, crawl events related web pages from the Internet set,
3. determining the event start time stamp \( t_s \) according to the web page set \( \varphi(t_i, t_j) \) and the event end time stamp \( t_e \) according to the download page time,
4. obtaining timing source data of event every day from the web page set \( \varphi(t_i, t_j) \) including new increasing web page number \( |\Delta \varphi(t_i, t_j)| \), new increasing keywords number \( |\Delta K(t_i, t_j)| \), and the distribution of event keywords in the new increased web page: \( \psi(t_i, t_j) \),
5. doing step (4) in the different sources of information (including news, blogs, and BBS).

We took “days” for the minimum time granularity, collected, from different sources on the web event, every day, the semantic characteristics of the source data, used the iterative algorithm to calculate the semantic outbreak powers every day based on the source data, and then got the events time-series data of the semantics outbreak over a period of time, as shown in Figures 3, 4, and 5.

5. Conclusions

Nowadays, emergencies events have a great impact on people’s daily lives. And, with the popularization and development of web technology worldwide, web is acting as a platform for the diffusion and evolution of social events. However, faced with the huge disorder and continuous web resources, it is impossible for people to efficiently recognize, collect, and organize the events. Therefore, automatically collecting and organizing the information about events and then tracking the evolution process of events becomes a hot research field. Generally, the traditional topic detection and tracking (TDT) techniques have been attempting to detect or cluster news stories into these events, without discussing or interpreting the evolution process of events.

In this paper, we proposed the semantic outbreak power and events process measurement algorithm. For any one event, we can calculate the event time-series data of the semantic outbreak power and measure the evolution process of events by analysing the information from the web about the event. It can help people to understand clearly the evolution process of a web event. Then, we propose an approach to distinguish the event type based on the fuzzy recognition.
Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Science and Technology Major Project under Grant 2013ZX01033002-003, in part by the National High Technology Research and Development Program of China (863 Program) under Grants 2012AA011504, 2013AA014601, and 2013AA014603, in part by National Key Technology Support Program under Grant 2012BAH107B01, in part by the National Science Foundation of China under Grants 61300202 and 61300028, in part by the Science Foundation of Shanghai under Grant 13ZR1452900, in part by the Major Research Project of the Ministry of the Public Security under Grant 2014JSYJ009, in part by the China Postdoctoral Science Foundation under Grant 2014M560085, in part by the China National Social Science Fund 06BFX051, and in part by the Shanghai University Training and Selection of Outstanding Young Teachers in Special Research Fund hzsf05046.

References

[1] J. Xuan, X. Luo, and J. Lu, “Mining websites preferences on web events in big data environment,” in Proceedings of the 16th International Conference on Computational Science and Engineering (CSE’13), pp. 1043–1050, 2013.

[2] 2012, http://definitions.uslegal.com/e/emergency-event.

[3] C. C. Yang, X. Shi, and C.-P. Wei, “Discovering event evolution graphs from news corpora,” IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans, vol. 39, no. 4, pp. 850–863, 2009.

[4] Y. Liu, L. Ni, and C. Hu, “A generalized probabilistic topology control for wireless sensor networks,” IEEE Journal on Selected Areas in Communications, vol. 30, no. 9, pp. 1780–1788, 2012.

[5] X. Luo, Z. Xu, J. Yu, and X. Chen, “Building association link network for semantic link on web resources,” IEEE Transactions on Automation Science and Engineering, vol. 8, no. 3, pp. 482–494, 2011.

[6] C. Hu, Z. Xu, Y. Liu, L. Mei, L. Chen, and X. Luo, “Semantic link network based model for organizing multimedia big data,” IEEE Transactions on Emerging Topics in Computing, vol. 2, no. 3, pp. 376–387, 2014.

[7] Y. Liu, Y. Zhu, L. Ni, and G. Xue, “A reliability-oriented transmission service in wireless sensor networks,” IEEE Transactions on Parallel and Distributed Systems, vol. 22, no. 12, pp. 2100–2107, 2011.

[8] Y. Liu, Q. Zhang, and L. Ni, “Opportunity-based topology control in wireless sensor networks,” IEEE Transactions on Parallel and Distributed Systems, vol. 21, no. 3, pp. 405–416, 2010.

[9] N. Y. Yen, T. K. Shih, and Q. Jin, “LONET: an interactive search network for intelligent lecture path generation,” ACM Transactions on Intelligent Systems and Technology, vol. 4, no. 2, article 30, 2013.

[10] R. W. H. Lau, N. Y. Yen, F. Li, and B. Wah, “Recent development in multimedia e-learning technologies,” World Wide Web, vol. 17, no. 2, pp. 189–198, 2014.

[11] Z. Xu, X. Wei, X. Luo et al., “Knowle: a semantic link network based system for organizing large scale online news events,” Future Generation Computer Systems, vol. 43–44, pp. 40–50, 2015.

[12] Z. Xu, X. Luo, S. Zhang, X. Wei, L. Mei, and C. Hu, “Mining temporal explicit and implicit semantic relations between entities using web search engines,” Future Generation Computer Systems, vol. 37, pp. 468–477, 2014.

[13] C.-C. Chen, T.-C. Huang, J. J. Park, H.-H. Tseng, and N. Y. Yen, “A smart assistant toward product-awareness shopping,” Personal and Ubiquitous Computing, vol. 18, no. 2, pp. 339–349, 2014.

[14] N. Y. Yen, C.-L. Wang, S. Hussain, and J. H. Park, “Computational awareness towards green environments,” The Journal of Supercomputing, vol. 69, no. 3, pp. 1007–1012, 2014.

[15] 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, pp. 194–218, 1998.

[16] J. Makkonen, “Investigation on event evolution in TDT,” in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language, pp. 43–48, 2003.
the 16th International Conference on Advanced Communication Technology (ICACT ’14), pp. 1242–1246, Pyeongchang, Republic of Korea, February 2014.

[27] Y. Xiaolin, Z. Xiao, K. Nan, and Z. Fengchao, “An improved Single-Pass clustering algorithm internet-oriented network topic detection,” in Proceedings of the 4th International Conference on Intelligent Control and Information Processing (ICICIPI ’13), pp. 560–564, Beijing, China, June 2013.

[28] R. Papka, On-line new event detection, clustering, and tracking [Ph.D. thesis], Department of Computer Science, University of Massachusetts, Boston, Mass, USA, 1999.

[29] L. Xiaowei, “Application of decision tree classification method based on information entropy to web marketing,” in Proceedings of the 6th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA ’14), pp. 121–127, IEEE, Zhangjiajie, China, January 2014.

[30] Y. Yang, T. Ault, T. Pierce, and C. W. Lattimer, “Improving text categorization methods for event tracking,” Data Mining and Knowledge Discovery, vol. 5, no. 3, pp. 167–175, 2004.

[31] X. Liu and I. Tang, “Mass classification in mammograms using selected geometry and texture features, and a new SVM-based feature selection method,” IEEE Systems Journal, vol. 8, no. 3, pp. 910–920, 2014.

[32] T. Leek and S. Sista, “Probabilistic approaches to topic detection and tracking,” Data Mining and Knowledge Discovery, vol. 7, no. 3, pp. 67–83, 2003.

[33] J. Allan, R. Papka, and V. Lavrenko, “On-line new event detection and tracking,” in Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 37–45, Melbourne, Australia, 1998.

[34] Y. Liu and X. Luo, “A framework for online hot event discovery on the web,” in Proceedings of the 16th IEEE International Conference on Computational Science and Engineering (CSE ’13), pp. 989–996, Sydney, Australia, December 2013.

[35] L. Deng and B. Xu, “Event evolution analysis in microblogging based on a view of public opinion field,” in Proceedings of the 6th International Symposium on Computational Intelligence and Design (ISCID ’13), pp. 193–197, 2013.

[36] P. Lee, L. V. S. Lakshmanan, and E. E. Milios, “Incremental cluster evolution tracking from highly dynamic network data,” in Proceedings of the 30th IEEE International Conference on Data Engineering (ICDE ’14), pp. 3–14, April 2014.

[37] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Information Processing and Management, vol. 24, no. 5, pp. 513–523, 1988.

[38] X. Yin, J. Han, and P. S. Yu, “Truth discovery with multiple conflicting information providers on the Web,” IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 6, pp. 796–808, 2008.