World Wide Web (2014) 17:377–404
DOI 10.1007/s11280-013-0207-7

Extracting news blog hot topics based on the W2T Methodology

Erzhong Zhou · Ning Zhong · Yuefeng Li

Received: 8 October 2012 / Revised: 22 February 2013 /
Accepted: 27 February 2013 / Published online: 26 March 2013
© Springer Science+Business Media New York 2013

Abstract Although topic detection and tracking techniques have made great progress, most of the researchers seldom pay more attention to the following two aspects. First, the construction of a topic model does not take the characteristics of different topics into consideration. Second, the factors that determine the formation and development of hot topics are not further analyzed. In order to correctly extract news blog hot topics, the paper views the above problems in a new perspective based on the W2T (Wisdom Web of Things) methodology, in which the characteristics of blog users, context of topic propagation and information granularity are investigated in a unified way. The motivations and features of blog users are first analyzed to understand the characteristics of news blog topics. Then the context of topic propagation is decomposed into the blog community, topic network and opinion network, respectively. Some important factors such as the user behavior pattern, opinion leader and network opinion are identified to track the development trends of news blog topics. Moreover, a blog hot topic detection algorithm is proposed, in which news blog hot topics are identified by measuring the duration, topic novelty, attention degree of users and topic growth. Experimental results show that the

E. Zhou · N. Zhong (✉)
International WIC Institute, Beijing University of Technology,
Beijing 100124, People's Republic of China
e-mail: zhong@maebashi-it.ac.jp

E. Zhou
e-mail: zez2008@emails.bjut.edu.cn

N. Zhong
Department of Life Science and Informatics, Maebashi Institute of Technology,
460-1 Kamisadori-Cho, Maebashi 371-0816, Japan

Y. Li
Faculty of Science and Technology, Queensland University of Technology,
Brisbane QLD 4001, Australia
e-mail: y2.li@qut.edu.au

proposed method is feasible and effective. These results are also useful for further studying the formation mechanism of opinion leaders in blogspace.

Keywords Wisdom web of things · Information granularity · Topic detection · Opinion leader · Topic hotness evaluation

1 Introduction

A blog is an online diary in the form of a Web page. As a product of Web 2.0, the blog provides information sharing and opinion interaction service. The blogspace is composed of blogs and related links. Nowadays, users more and more rely on the virtual blogspace to meet some needs in emotional expressions and information retrieval. Topic detection and opinion mining in blogspace are also valuable in information recommendation and business domain. A news blog is a kind of simple and timely news media. Topics in a news blog cover a wide range of life and are derived from news events such as social events, political events and economic events. A large number of Internet users are attracted by news blogs. Although topic detection and tracking techniques make great progress, researchers are still confronted with many problems. First, both information overload and personalized management cause big problems for topic mining and opinion mining in blogspace. Second, users do not make a uniform standard for the topic hotness. Third, what determines the formation and development of a blog hot topic is still an unsolved problem.

A news blog hot topic detection and tracking system can be a typical use case of the broad-based W2T (Wisdom Web of Things) methodology, which gives a perspective on how to unify humans models, information networks and granularity for analyzing the intersectional problems between the social world and the cyber world [55, 56]. The wisdom referred to W2T means that the service object is not isolated and the related context must be taken into account. Networks and information granularity referred in the W2T methodology provide a means for human centric computing. As for the idea of networks, the complex network theory emphasizes the structural change to analyze the related phenomenon or thing. For example, some measures proposed by the complex network theory are successfully used to discover the online community or study the information propagation mechanism in the community [5, 32]. On the other hand, the social network analysis theory emphasizes the relationship between individuals in a community and is widely used in sociological studies. For instance, the identification of user roles is often based on the social network analysis theory [59]. As for the idea of granularity, the granular computing theory provides a problem solving strategy by means of multiple levels of information granularity [48]. Moreover, granular computing is one of theoretical backgrounds of W2T methodology, which is used in information retrieval and Web usage mining. For example, the framework of information retrieval support systems (IRSS) that emphasizes the multiple representations of Web information is a typical application [47]. The document space, the query space, the term space and retrieval results are granulated in IRSS. A Web usage mining method that emphasizes the granular structure of user behavior data is another application [52]. In the method, the user behavior data in Web pages are first granulated according to the structure of

the website and service content and then the user motivation and potential intention are inferred. The W2T also lays emphasis on the influence and characteristics of humans regarding applications in various services. If the theory related to W2T is applied to an information recommendation system, unifying the ideas of humans, networks and information granularity means that user personality, context and need are the main inputs of the system. The mechanism of such a system can be explained as follows. First, the context is mapped into a network model. Then the structure of the network and relationships among entities within the model are analyzed, and the user need is analyzed on the basis of the characteristics of the network and user personality. Finally, different information granularities are identified according to the nature of the user need, and the related information is organized and processed for the user.

Previous studies have recognized that the user is the most important factor in Web information propagation. For example, the user thought, interaction pattern [15], participation degree of influential users [1, 39], user motivations [29], user backgrounds [31], and characteristics of neighbors [10] all influence the information propagation. At the same time, the problem solving strategy of W2T lays emphasis on human nature. Hence, in order to detect hot topics in news blogs, the paper is based on the W2T methodology to analyze the formation and development of a blog topic. Namely, the context of topic propagation is first decomposed into the different categories of complex networks related to users. Then the influence of different kinds of users and network opinions on a topic is measured in related networks corresponding to the special information granularity. Finally, the paper concludes that the user motivation and behavior pattern determine the burst and temporal features of a news blog topic. Moreover, the user behavior pattern, network opinion and opinion leader play a vital role in different phases of a blog topic.

The rest of the paper is organized as follows. Section 2 introduces the related work on hot topic detection. Section 3 analyzes the related problems. Section 4 describes the proposed method. Section 5 tests the feasibility and effectiveness of the proposed method by some experiments. Section 6 gives conclusions and future work.

2 Related work

In the Topic Detection and Tracking (TDT) domain, a topic is defined as a seminal event or activity, along with all directly related events and activities [12]. As far as topic detection techniques are concerned, vector space model [12, 17, 41, 43, 57], probabilistic models [7, 9], and complex network theory methods [36, 54, 58] are popular. He and colleagues used incremental term frequency inverse document frequency model and incremental clustering algorithm to detect new events [17]. Chen and colleagues first extracted the hot words based on the distribution over time and life cycle, then identified key sentences and grouped the key sentences into the clusters that represent hot topics [12]. Zhou and colleagues adopted Density-based Spatial Clustering of Application with Noise (DBSCAN) method to group words into word clusters so as to extract popular topics [57]. Wang and colleagues used Unweighted Pair Group Method Using Arithmetic Averages (UPGMA) algorithm to detect news topics [41]. Wang and colleagues proposed an improved k-means

method to detect hot topics [43]. Brants and colleagues used Probabilistic Latent Semantic Analysis (PLSA) model to detect different topics in documents [9]. Blei and colleagues analyzed topics in documents by Latent Dirichlet Allocation (LDA) method [7]. Zhu and colleagues constructed the word network based on word co-occurrence and extracted keywords from the texts according to the small world structure of the word network [58]. Shi and colleagues analyzed the topic of the text according to the small world structure of text words [36]. Zhao and colleagues extracted keywords from the document according to complex network theory and the small world structure of words. The method based on the vector space model is widely used for the simplicity [54]. However, the vector space model lacks semantic correlation and has the high dimension problem of features. PLSA and LDA models rely on a large number of sample data. The method based on complex network theory is more complicated than other methods.

A hot topic is the topic which users widely discuss and consecutively pay attention to. He and colleagues considered the frequency and consecutive time of news reports to evaluate the topic hotness [17]. Gong measured the topic hotness from user participation and media report [16]. Li and colleagues realized the blog topic hotness evaluation by combining the number of reviews, comments, opinions, and publication time [24]. As for the topic hotness evaluation, most of the evaluation strategies mentioned above, mainly focus on the reactions of users without further considering the nature of topic hotness. Namely, the hotness of a blog topic reflects the interest degree of all users, and the changes of the topic hotness mean that the interests of users are influenced by some factors. Hence, the hotness evaluation cannot ignore the changes of the factors that influence the development trends of topics.

In general, a hot topic often emerges when the isolated individual behavior develops into the mass behavior. However, it still is a complicated problem to answer why a blog topic grows into a hot one, because the topic hotness is related to many factors such as the event, Web media and users. The research strategy in the paper is different from the traditional strategies, because we are aware of the following phenomena. First, users both attach importance to the development of a news event and pay attention to the opinions published by other users. A network opinion is consequently considered to be an important factor for the growth of a topic, and the changes of network opinions on the topic need to be detected. Second, the influence of opinion leaders can be enlarged by the Web so that the development of the new event is sometimes determined by the opinion leaders. Hence, the features and formation conditions of opinion leaders are carefully analyzed, and the influence of the opinion leader on the topic is measured.

Although researchers have realized that the evolution of an online community mainly arises from the changes of user behavior [23], they did not emphasize the reason for influencing user decisions. That is also the reason why the W2T methodology lays emphasis on understanding the phenomena in the cyber world from the perspective on humans. Recently, many scholars reconsidered the traditional problems in the light of the system framework or methodology of the W2T. The evolution of online opinions and the analysis of user online behavior have been crucial issues in the applications of Web intelligence. For example, Liu and colleagues attempted to integrate user opinions, product sales and computer systems into an entity, which represents a direct application of the W2T data cycle system [27]. Msuical and colleagues adopted the method of complex network theory to make an

investigation on user behavior and interactions in the online world for personalized services in the W2T [30].

3 Problem analysis

In order to construct an effective topic model to represent news blog topics and reasonably evaluate the topics, this section investigates two problems ignored by the traditional methods. First, what are the characteristics of news blog and blog users, and how an opinion leader emerges as a special blog user. Second, how to represent and decompose the context of topic propagation into related complex networks, and how to identify the key factors that determine the development trends of news blog topics in different complex networks.

3.1 Characteristics of a news blog

The development of a topic is related to the characteristics of the related blog. Generally speaking, blogs can be categorized into three types according to the contents of these posts, namely professional technique, individual life and temporal topics [34]. Topics in the professional blog mainly refer to professional techniques related information, but topics in the life blog are with respect to individual life affairs. Moreover, topics in the temporal blog are mainly related to news events. A news blog inherits the characteristics of the temporal topic blog. Compared with the professional blog topics, topics in the news blog often show burst and temporal features. Users visit different kinds of blogs with different motivations [34]. For example, users share the specific knowledge and learn from each other in professional blogs. Users construct social networks or maintain their social contacts by means of individual life blogs. However, bloggers publish posts in news blogs with the purpose of revealing the truth of a news event or drawing other users' attention to the special event. Users with their own different backgrounds take part in topic interaction for the sake of knowing the truth of a news event or expressing their own personal feeling. The relationships among members in a topic group are weak because the topic group is mainly maintained by user interests [40]. The topic group dissolves when users lose interests in the related topic. Hence, the user motivation determines the burst and temporal features of a news blog topic to a great extent.

3.2 Development of a blog topic

Topics usually can be categorized into three types, namely stable, transient and fluctuant ones. A *stable* topic can last a period of time, but few users pay attention to it. A *transient* topic often rapidly vanishes on the information ocean. A *fluctuant* topic is likely to become a hot one, and the paper attaches importance to the topic with fluctuant features. According to previous studies of journalism, such a topic can experience its own life cycle with the birth, growth, maturity and fade phases [50]. At the beginning, one blogger sponsors an issue, and other bloggers are attracted to join. When the topic enters into the growth phase, users actively express opinions or spread the topic because the news event may be obscure and Web information is not credible in the early phase. While the topic group rapidly grows up, opinion leaders

emerge. The opinion leaders hold the most representative network opinions and are influential in giving birth to the public opinion. If the topic is disputable, the group is often separated into the cliques that support different opinions. The more attractive the topic is, the more active the opinion leaders are. When the topic grows up into the maturity, the public opinion emerges. If the majority of users are not satisfied with the disposal of an event, the public opinion may evolve into other opinions. At last, the influence of an opinion leader gradually becomes weak and then the topic fades away.

According to the description mentioned above, it is undoubted that the user, topic and opinion are correlated with each other in blogspace. Considering the perspective on W2T methodology, the human is equivalent to a blog user in blogspace. The network is equivalent to the context of topic propagation. Information granularity presents the semantic level of information expression. In order to analyze the formation and development of a blog topic, three complex networks are extracted from the blogspace. As shown in Figure 1, the blog community, topic network and opinion network are the main contexts of topic propagation. The blog community is composed of users with similar interests. The topic network is composed of topics with different granularities and evolutionary relationships among topics. The opinion network is composed of opinions and interaction relationships among users.

The boundary of a blog community can be identified by blog topics. Moreover, the evolution of network opinions is correlated with the evolution of a topic. Hence, the construction of a topic network is the key to analyzing the related phenomena during the topic propagation. With the development of a news event, the structure of the topic network changes accordingly. In order to represent the changes of blog topics from both microcosmic and macroscopic aspects, the information granularity is

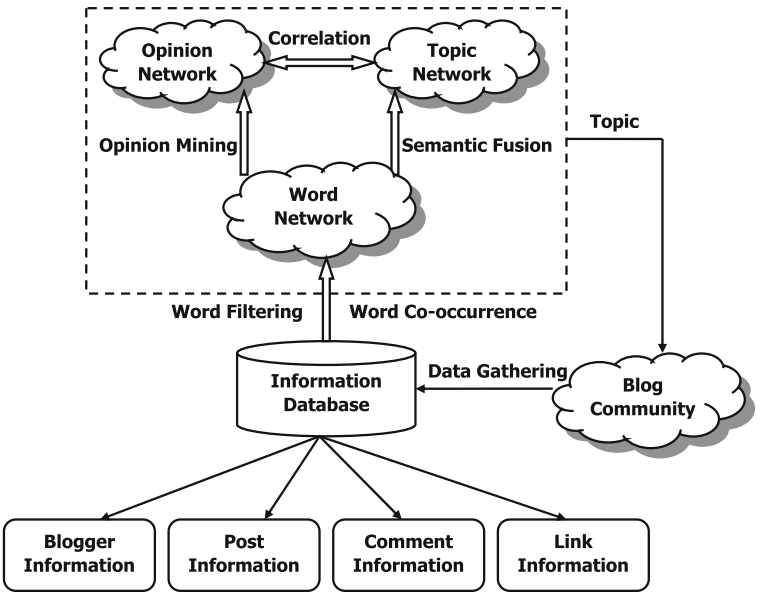

Figure 1 Complex networks in blogspace

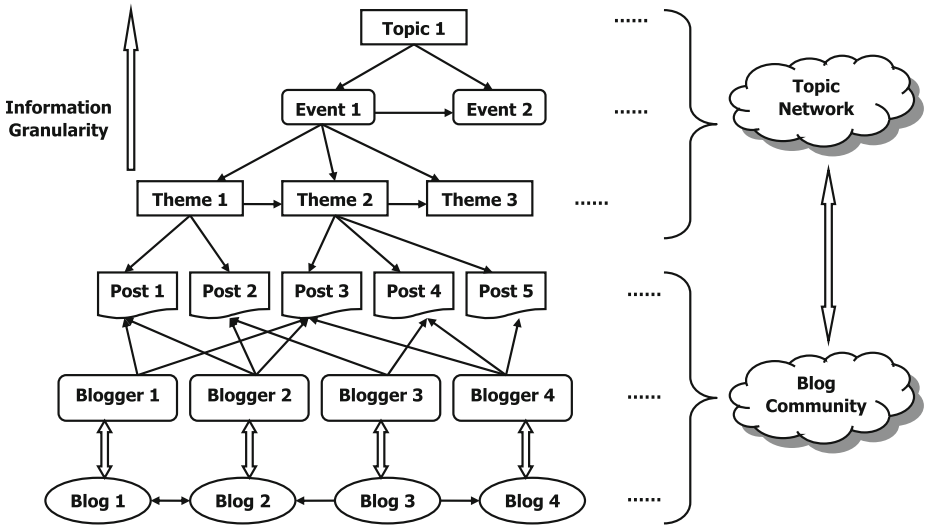

Figure 2 The structure of a topic network in blogspace

taken into account. The hierarchical structure of a topic network is shown in Figure 2. The information in different layers shows the characteristics of human information processing and starting points of topic propagation.

The blog community has different structural characteristics within different life phases. According to social network analysis theory, the structure of a community is related to interaction patterns and relationships among members. Figures 3 and 4 show that the comment trends of blog hot topics from the China Sina blog website in 2011. Topics in Figure 3 are derived from domestic events in China, and topics in Figure 4 are opposite. The obvious phenomena are that most of the users are active in the early phase. The phenomena can be explained as follows. Users lack authoritative information and are very curious about the event in the early phase. Hence, the effect of sheep flock occurs in the community. When a public opinion

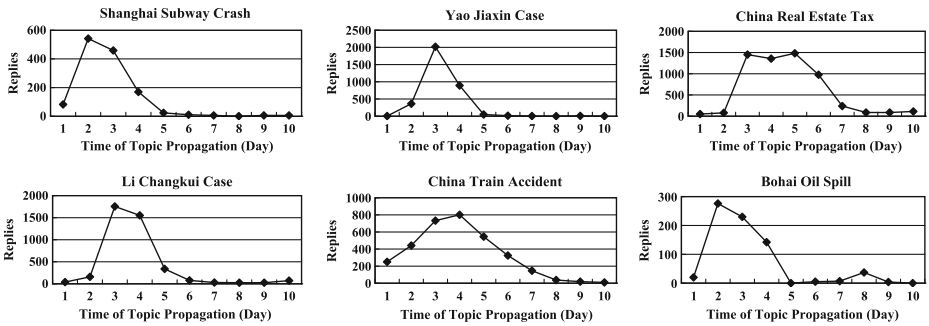

Figure 3 The comment trends of the hot topics related to news events in China

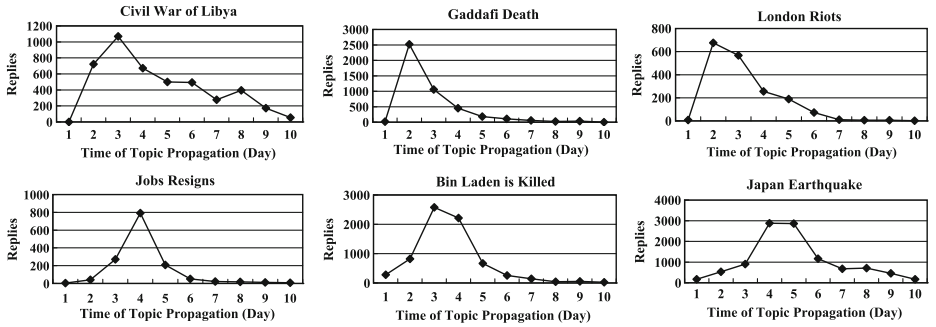

Figure 4 The comment trends of the hot topics related to news events in other countries

comes into being, the spiral of silence emerges. Namely, users often keep silent when their opinions are in the minority. Moreover, the novelty of topics gets weaker and weaker with the lapse of time so that the behavior patterns of users have drastic changes. The structural change of a blog community can consequently reflect the development trend of a topic, and the behavior patterns of users also determine the burst and temporal features of blog topics. As far as the user role is concerned, users evolve into three categories, namely opinion leaders, active users and ordinary users. The opinion leader is influential during the evolution of a network opinion. An active user plays the role of a disseminator. However, the authority of such an active user is often weak in comparison with the opinion leader. Ordinary users often lurk after the public opinion emerges. As for the vocational background, a specialist often publishes more posts than an ordinary user and can easily become an opinion leader. Moreover, the specialists often publish posts on weekday, and some specialists even regard the writing as a career. On the contrary, the ordinary users publish posts on weekend with the purpose of a pastime [53]. Based on the analysis mentioned above, it is clear that the user behavior pattern, network opinion and opinion leader all impact on the development of a blog topic.

The blog community represents the relationships among users, and the opinion network represents the construction process of those relationships. As for a hot topic, network opinions on a topic often experience great changes through the whole life span. Hence, the structural changes of an opinion network can represent the opinion interaction degree of users and the evolutionary trend of network opinions. Moreover, opinion leaders can also be recognized by analyzing the characteristics of such an opinion network. For example, the structure and sentimental polarity distribution of an opinion network can contribute to recognizing the opinion leader during the topic propagation [8]. At the same time, the sentimental polarity distribution of an opinion network also reflects the development trends of network opinions.

3.3 What is an opinion leader?

A typical phenomenon during the evolution of network opinions is the emergence of opinion leaders. An opinion leader is the blog user who plays a crucial role in influencing opinion makings of other users. However, opinion leaders in blogspace

are different from those in a physical community. The high popularity and recognition are the basic conditions of opinion leaders. At the same time, the ordinary users do not pay more attention to the opinion leader's social background in the physical world. Moreover, the diffusion speed of Web accelerates the formation of an opinion leader in blogspace. The burst feature of opinion leaders is very obvious in blogspace. According to the previous studies, the opinion leaders can be categorized into three types, namely fluctuant, long-term and transient opinion leaders [44, 49]. The fluctuant leader often changes his/her own personal role during the evolution of a network opinion. The long-term leader is active to participate in topic interaction in order to gain the better recognition and popularity. The transient opinion leaders often either voluntarily keep silent after they publish a special idea or passively lose the influence because of the opposition from surroundings. The publication number of high quality posts is a main difference between the long-term leader and the transient leader. The fluctuant leader is not interested in promoting the recognition and popularity in comparison with the long-term leader. For example, some public figures often take part in opinion interaction for controlling the moods of other users when network opinions are in issue.

It is an important task for information recommendation and opinion monitoring to identify opinion leaders in blogspace. As for the identification of user roles, the influence of a user on neighbors or other users in the community is often measured. Song and colleagues evaluated user influence by measuring the importance and novelty of posts [39]. Bodendorf and colleagues detected opinion leaders by means of social network analysis [8]. Akritidis and colleagues evaluated user influence from the temporal aspect of user behavior and the quality of related posts [2]. Lim and colleagues first constructed a social network according to users' activities and usage patterns and then identified the influential users by evaluating the quality of posts in the social network [26].

4 Detecting hot topics in news blogs

This section presents a topic detection approach that takes into account the different views of event reports and evolutionary relationship between events. The information granularity is also considered in the construction of a topic model. In order to detect the current and forthcoming hot topics, the growth state of each topic is considered in addition to user interests because the growth state not only indicates the development trend of the topic but also represents the vitality of topics. The hotness of a topic is evaluated by measuring the duration, topic novelty, attention degree of users and topic growth.

4.1 Topic model and algorithm for detecting hot topics

A blog topic is a kind of Web information and the Web document is an important information carrier. In Web content mining, information granularity needs to be carefully considered in information processing such as information extraction and information arrangement for Web documents. As for the structure of the body of a Web document, the section, paragraph, sentence and word are oriented to different information granularities. Hence, the clarity of Web information is closely related to

information granularity. As for information arrangement, documents can be grouped into the different scales of clusters according to categories of topics. Moreover, if each section in a document discusses a specific issue or subtopic, the topic can be divided into issues or subtopics. As for the user expression, users can comment on one topic from different aspects and publish posts with different writing styles because of the personalized management. Hence, the structure of a blog topic can be characterized by multi-level or multi-view.

Considering information granularity in the semantic expression, a topic can be considered to be a cluster that is comprised of related events, and an event can be considered to be a cluster that is comprised of related themes. As for the literal expression of a theme, the theme can be expressed by a group of related keywords. Hence, the topic can be divided into three layers. The models in three layers are as follows:

- A topic model is defined as $Topic = \{Event_1, Event_2, \ldots, Event_n\}$, where $Event_i$ denotes the ith event;
- An event model is defined as $Event = \{Theme_1, Theme_2, \ldots, Theme_m\}$, where $Theme_i$ denotes the ith theme;
- A theme model is defined as $Theme = \{Keyword_1, Keyword_2, \ldots, Keyword_s\}$, where $Keyword_i$ denotes the ith keyword of a post.

It is very difficult to measure the information granularity due to the complicated semantic expressions, and sometimes it is not easy to distinguish the information in the topic layer from the one in the event layer. Hence, it is essential to analyze the difference between a topic and an event. News blog topics are often derived from an event or activity. As for a natural disaster, the keywords in event reports are involved in the situation of the disaster in the early phase. Then some of the keywords shift to the disposal and precaution of the disaster. The event related to the disaster often cannot evolve into other events because of the strict manual control, and the topic model is actually oriented to the whole event. For example, the construction of the topic model related to an accident is shown in Figure 5. The arrow points the next state of a model, and the development of such a topic is reflected in three layers. The evolution of theme models reflects the shift of user issue on the subway accident, and the same event model or topic model changes accordingly. However, the activity is composed of a series of related events such as the president campaign and Olympics. As for the topic related to Olympics, users attach importance to different events

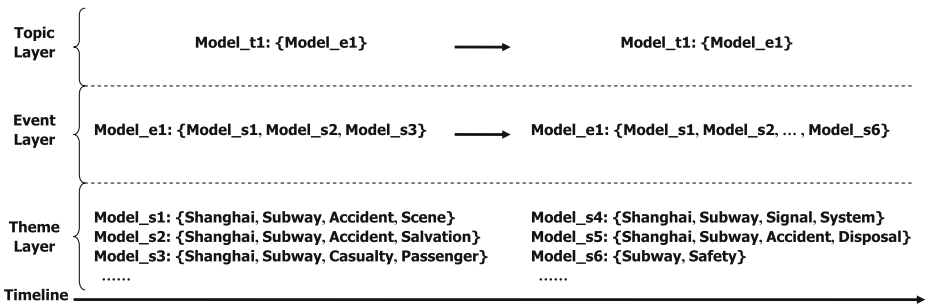

Figure 5 The construction of the topic model related to a subway accident

with the lapse of time. The preparations for Olympics, different kinds of games and performance of the national team in Olympics are reported one after another. Moreover, the topics oriented to the activities are complicated, and the evolutionary relationship between events needs to be carefully considered.

The hot topic detection algorithm presented in Algorithm 1 can be divided into three parts, namely the topic detection (Steps 1−13), opinion leader identification (Steps 14−17) and hotness evaluation (the rest of Algorithm 1), respectively. Topic detection task focuses on the posts published within different time intervals due to the temporal feature of a news topic. Hence, the time complexity of topic model construction that adopts post clustering is $O(n)$. The emergence of an opinion leader means that the influence of a blogger reaches a specific level by means of the accumulation of popularity and recognition during a period of time. Hence, the worst time complexity of opinion leader identification task is $O(n^2)$. The topic hotness evaluation task focuses on the topics that are active at a specific time interval and its time complexity is consequently $O(n)$. Algorithm 1 is composed of above three tasks and its time complexity is $O(n^2)$. The following three subsections will give the descriptions of the three parts in details.

Algorithm 1 Hot Topic Detection

Input: S is a post set, n is the number of time units, d is the threshold of hotness.
Output: T_{set} is a hot topic set.

1. **for** each post i in S **do**
2. Extract keywords from i;
3. Construct the theme model of i;
4. **end for**
5. **for** time unit $j = 0$ to n **do**
6. Identify the posts related to the event according to correlated keywords;
7. Construct the event models based on the strategy of single pass clustering;
8. Add the event models into event list EL_j within j;
9. **if** j is equal to 1 **then**
10. Construct the topic models within j manually;
11. **else**
12. Construct a new topic model or update the previous topic model according to the related event model in EL_j;
13. **end if**
14. Analyze the structure of the blog community based on interaction relationships among users;
15. Detect network opinions on each topic according to the topic models;
16. Construct the opinion network within j on the basis of the blog community;
17. Identify opinion leaders with respect to each topic;
18. Count the changes of repliers, opinion leaders, replies and network opinions to measure the growth state of each topic within j;
19. Evaluate all topics within j by measuring the duration, topic novelty, attention degree of users and topic growth;
20. Select the topics whose hotness is more than d, and add new hot topics are into T_{set};
21. **end for**
22. Return T_{set}.

4.2 Topic detection based on the view of event reports

The event model is the key to constructing the topic model. News topic detection methods often rely on some studies of event detection and tracking [33, 45]. Two important tasks for constructing a blog topic model are listed. One is how to group the related posts that convey the information of different granularities to extract a news event. The other is how to track a news event by identifying the evolutionary relationship between events. The single pass clustering algorithm is widely used in event detection and tracking [3, 11]. The topic detection based on the view of event reports first adopts the strategy of single pass clustering to extract the events which occur in a given time interval. Then the evolutionary relationship between events is identified by computing the content similarity between events and the distribution of each event on different topics. At last, the topic model is created or updated by detecting the new event and tracking the previous event.

The topic detection based on the view of event reports copes with the posts that are in the chronological order and refines different topic models. The different strategies are adopted to deal with the information in different layers during the process of topic detection. In other words, the theme, event and topic models are constructed according to the different strategies. The following three sections will describe how to construct the three kinds of models. Section 4.2.1 points out how to pick keywords from a post in order to represent the post in the form of the theme model. The single pass clustering algorithm clusters the posts which may belong to the same event. Hence, the post clustering means the theme models are compared and then are clustered for constructing the event model. Section 4.2.2 describes the evaluation measure for the comparison between theme models. The identification of the evolutionary relationship between events determines the construction of a topic model. Section 4.2.3 introduces how to identify the temporal relationship between the neighbor events so as to create or update the topic model.

4.2.1 Keyword extraction and keyword correlation

A keyword is a basic element for the topic model. Nouns and verbs are selected from the title and the first paragraph of a blog post. The tags of a post are also picked as candidate words. At last, the keyword is identified according to the weight of the candidate word. The weight of the candidate word is evaluated by the following equation in the Term Frequency Inverse Document Frequency (TFIDF) method [35]:

$$Weight(t_k, r) = TF(t_k, r) * \log\left(\frac{N}{N_k + 0.5}\right) \tag{1}$$

where r is a blog post, t_k is a word, $Weight(t_k, r)$ is the weight of t_k in r, $TF(t_k, r)$ is the frequency of t_k in the text of r, N is the total number of blog posts, and N_k is the number of blog posts where t_k appears.

In order to increase the precision of the clustering algorithm, an information retrieval strategy is adopted to filter irrelevant posts in advance. Namely, the posts related to a special event are retrieved by the correlated keywords which can describe the event. If users frequently publish and comment on an event, some keywords are highly correlated within a given time interval. The chi-square test that

is successfully used to measure the correlation between keywords is presented as follows [6]:

$$\chi^2 = \frac{(E(uv) - A(uv))^2}{E(uv)} + \frac{(E(\bar{u}v) - A(\bar{u}v))^2}{E(\bar{u}v)} + \frac{(E(u\bar{v}) - A(u\bar{v}))^2}{E(u\bar{v})} + \frac{(E(\bar{u}\bar{v}) - A(\bar{u}\bar{v}))^2}{E(\bar{u}\bar{v})} \quad (2)$$

$$E(uv) = \frac{A(u) * A(v)}{N} \quad (3)$$

where $A(u)$ is the number of the posts which word u appears in, $A(\bar{u})$ is the number of the posts which word u does not appear in, $A(uv)$ is the number of the posts which both words u and v appear in, and N is the total number of posts. If a keyword set includes at least two keywords, each keyword pair is extracted at random for the post retrieval.

4.2.2 Post clustering

The single pass clustering algorithm can be explained as follows. A report is merged into an event if the content similarity between two reports is above a threshold. Otherwise, the report is considered to be the first report of a new event. When the text is represented in a word set, the Jaccard coefficient which computes the overlap ratio between two sets is often used to evaluate the similarity between two texts [19, 42]. The larger the value of the Jaccard coefficient is, the higher the similarity between two texts is. However, as shown in Figure 6, the Jaccard coefficient does not consider the size of a set, and the effect of text clustering is not good when two sets are not at the same level. In our method, a post is represented in its own theme model, and the similarity evaluation between posts is based on the similarity between the theme models. Hence, the proposed method compares the difference in size between two sets on the basis of the Jaccard coefficient, and the following equation is used to compare the similarity between two posts:

$$Sim(d_i, d_j) = \alpha * \frac{|d_i \cap d_j|}{|d_i \cup d_j|} + \beta * \frac{|d_i \cap d_j|}{min(|d_i|, |d_j|)} \quad (4)$$

where d_j denotes the keyword set of the jth theme, $Sim(d_i, d_j)$ is the similarity between the ith theme and the jth theme, $d_i \cap d_j$ is the intersection of sets d_i and d_j,

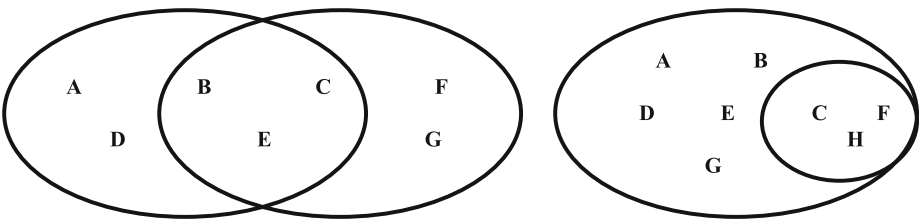

Figure 6 The intersection of two sets

$d_i \cup d_j$ is the union of sets d_i and d_j, $|d_i|$ is the size of set d_i, α and β are coefficients, and $min(y, z)$ is the minimum between y and z.

4.2.3 Identification of the evolutionary relationship between events

Figure 7 shows the two cases of event evolution, the left case shows that one event may evolve into other two events, and the right case shows that multiple events may evolve into a new event. The attributes of events sometimes change a lot so that the evolutionary relationship between events needs to be carefully considered. As for the work of Hong and colleagues, the topic and report are divided into some subtopics and then the new topic is detected according to the proportion and distribution of relevant subtopics [18]. If the proportion of relevant subtopics is low and the distribution of such subtopics is dispersive, the report is classified into a new topic. On the other hand, when the relationship between events is evaluated, the time when the event happens is the other key factor in addition to the content similarity between events. If two events occur very closely, the events are likely to be correlated with each other. On the contrary, the events are unlikely to be correlated even if the content similarity between them is very high. The proposed method adopts Hong's strategy. The topics that are relevant to the target event within the adjacent time interval are identified according to the content similarity between events and then the evolutionary relationship between events is identified by means of the distribution of the relevant topics. The evolutionary relationship between two events is identified by the following process:

1. The topic model $Topic_k$ which the original event $Event_i$ belongs to is identified;
2. The similarity event set $S_{ij}(i \neq j)$ within the former time unit is constructed for the target event $Event_j(j \neq i)$ according to the content similarity between events;
3. If S_{ij} is not empty and most of the events in S_{ij} belong to $Topic_k$, there exists an evolutionary relationship between $Event_i$ and $Event_j$;
4. If S_{ij} is empty, the similarity event set S'_{ij} in which similar events for $Event_j$ occur within the current time unit and are earlier than $Event_j$ is constructed;
5. The same strategy stated above for S'_{ij} is adopted.

If there is only one similar event for the target event and the content similarity is very high, the two events are considered to be the same event.

According to the definition of the event model, the theme model is the basis of the event model. Hence, the comparison of content similarity between events means that the related theme models need comparing. When two events are compared, each

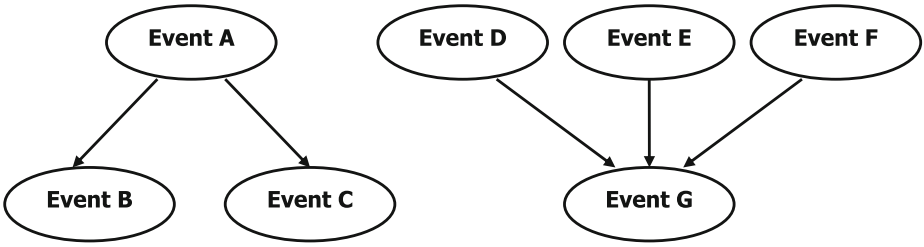

Figure 7 The evolutionary relationship between events in different cases

theme in one event model is compared with that in the other event model, and the following event similarity equation is used to measure the content similarity between two events:

$$Comp(e_i, e_j) = \frac{1}{cn * cm} \sum_{p=1}^{cn} \sum_{q=1}^{cm} Sim(d_{ip}, d_{jq}) \tag{5}$$

where $Comp(e_i, e_j)$ denotes the similarity between events e_i and e_j, cn is the number of the themes that belong to event e_i, cm is the number of the themes that belong to event e_j, d_{ip} denotes the pth theme of event e_i, and d_{jq} denotes the qth theme of event e_j.

4.3 Identification of opinion leaders

The identification of opinion leaders is important in evaluating the development trend of a blog topic. An opinion leader is characterized by high popularity and high recognition. As far as the structure of the social network is concerned, opinion leaders have a larger central degree than other users because the leaders have a group of followers. Moreover, they have a profound insight into a social phenomenon so that their opinions can become representative opinions.

The influence of a user on network opinions can be measured by counting the sentiment polarity distribution of neighbors' opinions. Hence, an opinion network is useful to identify opinion leaders. The blog community based on user comments or post quotation is an important precondition for constructing such an opinion network. Moreover, user opinions are extracted on the basis of a topic model, and the sentiment of each opinion is analyzed. Based on those preparations, the opinion networks within different time intervals are constructed. As shown in Figure 8, nodes stand for users' opinions and an edge between nodes stands for the interaction relationship between users. Moreover, the black node stands for a positive opinion, and the white node stands for a negative opinion. The arrow of each edge points to the commented object.

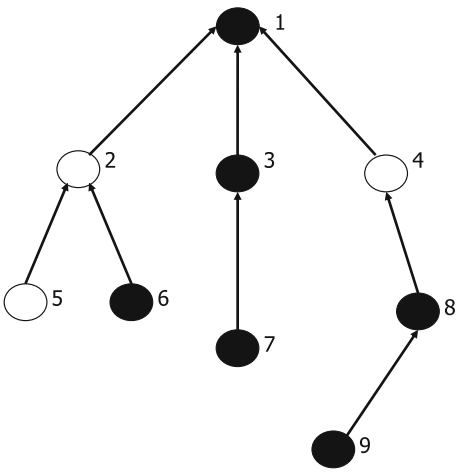

Figure 8 The opinion network based on the opinion interaction relationship

4.3.1 Opinion extraction and sentiment analysis

Opinion mining is important for opinion leader identification and is composed of the topic extraction, opinion holder identification, claim selection and sentiment analysis. The topic extraction lays emphasis on identifying the commented objects. However, the commented objects are oriented to different event attributes, and sometimes one user may criticize the other user. Hence, the topic extraction needs to take the context of opinion interaction into consideration to avoid the topic drift. Moreover, a user reply is arbitrary. For example, there are an intact sentence, sentence without subject and emotional word. Hence, the recognition of interaction relationship is the key to identifying the commented object when the user reply is a sentence without a subject or emotional word.

In general, the noun, verb, adjective and adverb are useful for the sentiment analysis of network opinions. The stop words in a sentence are consequently filtered to assure the accuracy of the sentiment analysis. The equation proposed by Ku et al. [22] for recognizing the sentiment polarity is presented as follows:

$$S(T) = \sum_{i=1}^{N_{\text{count}}} S_{ci} \tag{6}$$

$$S_{ci} = P_{ci} - N_{ci} \tag{7}$$

$$P_{ci} = \frac{\frac{fp_{ci}}{\sum_{j=1}^{n} fp_{cj}}}{\frac{fp_{ci}}{\sum_{j=1}^{n} fp_{cj}} + \frac{fn_{ci}}{\sum_{j=1}^{m} fn_{cj}}} \tag{8}$$

$$N_{ci} = \frac{\frac{fn_{ci}}{\sum_{j=1}^{m} fn_{cj}}}{\frac{fp_{ci}}{\sum_{j=1}^{n} fp_{cj}} + \frac{fn_{ci}}{\sum_{j=1}^{m} fn_{cj}}} \tag{9}$$

where $S(T)$ represents the sentiment polarity of sentence T, N_{count} is the number of the words that do not belong to stop words in sentence T, S_{ci} is the sentiment polarity of the ith word in sentence T, fp_{ci} represents the frequency of the ith word of sentence T in the positive opinion sample set, fn_{ci} represents the frequency of the ith word of sentence T in the negative opinion sample set, n is the total number of different words in the positive opinion sample set, m is the total number of different words in the negative opinion sample set, j represents the jth word in the opinion sample set, fp_{cj} is the frequency of the jth word in the positive opinion sample set, and fn_{cj} is the frequency of the jth word in the negative opinion sample set.

4.3.2 Post ranking

The post quality is often estimated according to the link relationship between posts. However, there are different kinds of links, and bloggers set different links for different purposes, such as social contact or information recommendation. Hence, the topic drift often has a negative effect on the post ranking algorithm if the link information is not analyzed. It is proven that the quality of a post is likely to be

high if the post has a long text, a large number of comments or in-links and much less out-links [1]. The paper consequently analyzes the categories of links and assigns different weights to the different kinds of links. The links that link a post to a blog are distinguished from those between posts. The hyperlinks between posts are filtered according to the content. The post influence equation which is based on Agarwal's method is described as follows:

$$I(p_a) = w_{\text{len}} * Len(p_a) * \left(w_{\text{com}} * Rp(p_a) + w_{\text{qu}} * Tr(p_a) + w_{\text{in}} * \sum_{i=1}^{m} I(p_i) - w_{\text{out}} * \sum_{j=1}^{n} I(p_j) \right) \tag{10}$$

where $I(p_a)$ is the influence of post p_a, $Len(p_a)$ is the length of post p_a, $Rp(p_a)$ is the number of the replies of post p_a, $Tr(p_a)$ is the number of the quotation of post p_a, w_{len} is the post length coefficient, w_{com} is the comment coefficient, w_{qu} is the quotation coefficient, w_{in} is the in-link coefficient, w_{out} is the out-link coefficient, i denotes the ith post that links to p_a, j denotes the jth post which p_a links to, m is the total number of the in-links of post p_a, and n is the total number of the out-links of post p_a.

4.3.3 User influence evaluation

The work of Li and colleagues shows that an influential blogger can be discovered by evaluating the user influence which is categorized into the social contact, the content of a post and activeness [25]. The proposed method is based on Li's evaluation strategy and opinion leaders are identified from those influential bloggers by evaluating their influence on online opinions. The quotation is the main representation of user influence in an online community. Hence, an information propagation network based on the quotation between posts is constructed, and the central degree of a node that stands for a blogger in an online community is counted. The larger the quotation number of a blogger is, the higher the user influence of social contact is. At the same time, opinion leaders often lead the direction of the consensus by means of related posts. However, browsers are not capable of reading full posts due to information overload. Hence, bloggers must try their best to publish more high quality posts in order to influence the more users. The user influence of the content of a post is assessed by both the ratio of user high quality posts to the total posts and the proportion of supporters in the topic group. As for the user influence of activeness, it is useful to promote the influence by taking part in the topic discussion. Hence, the user influence of activeness is evaluated by counting the number of user replies that are published during the opinion interaction and represent user opinions. The influence evaluation equation for identifying the opinion leader is defined as follows:

$$Opind_n(b, x) = \sum_{i=1}^{n} \frac{nop_i(b, x)}{Tnop_i(x)} * \left(\psi * cen_i(b) + \varphi * \frac{nos_i(b, x)}{Tnos_i(x)} + \delta * \frac{bp_i(b, x)}{Tp_i(x)} \right) \tag{11}$$

where $Opind_n(b, x)$ is the influence of user b with respect to topic x within the nth time unit, i denotes the ith time unit, $nop_i(b, x)$ is the number of the opinions that

user b publishes with respect to topic x within the ith time unit, $Tnop_i(x)$ is the total number of the opinions that all users publish with respect to topic x within the ith time unit, $cen_i(b)$ is the central degree of user b in the social network within the ith time unit, $nos_i(b, x)$ is the number of the users who have the same opinion as user b within the ith time unit, $Tnos_i(x)$ is the total number of the users who publish their opinions on topic x within the ith time unit, $bp_i(b, x)$ is the number of the high quality posts that user b publishes with respect to topic x within the ith time unit, $Tp_i(x)$ is the total number of the posts that all users publish with respect to topic x within the ith time unit, ψ is the location coefficient, φ is the emotion coefficient, and δ is the quotation coefficient.

4.4 Topic hotness evaluation

The characteristics of a news blog make it clear that the user interests play an important role. Moreover, the vitality of a topic is dynamic and the lifetime of a temporal topic is especially short. Hence, the interest degree of users and the growth degree of a topic are measured for detecting a hot topic.

As far as user interests are concerned, the context of topic propagation can stimulate users to participate in the topic interaction. If a topic spreads a long time, the topic can have a great probability of attracting users. If a topic is very new, the blog user is more likely to be interested in the topic. If the posts related to a topic have high quotation or a great number of replies, the topic is easily recommended to all users by the website. Hence, if a topic meets all the conditions mentioned above, user interests in the topic are undoubtedly high. User interests in a topic are consequently evaluated by measuring the duration of the topic, the topic novelty and the attention degree of users. The proposed method assesses the novelty of a topic in the light of human forgetting factors [4]. Namely, humans can clearly remember an event if the event happens recently or occurs frequently. Inversely, human memory can decay if the event happens long ago and seldom occurs. The degree of user attention is evaluated by counting the total number of post quotation and replies that a topic owns.

In order to detect the change of topical vitality, the topic growth needs to be evaluated on schedule. According to the analysis on the development of a topic in Section 3.2, the emergence of the online consensus means that a blog topic enters into the maturity phase that is an important turning point. The related research of communications finds that the online consensus can come into being if the following conditions are satisfied [28]. First, the number of replies reaches a specific level. Second, opinion leaders actively play their roles. Third, users' moods are ignited. Fourth, a great number of media are involved in the topic propagation. As for the work of Zhang and colleagues, the influence of the online consensus on an online topic is ignored and the topic growth is evaluated by the number of clicks and replies [51]. Although Zhang's measure strategy is adopted, the topic growth in the proposed method is evaluated according to the changes of replies, repliers, opinion leaders and online opinions within different time intervals. When opinions are counted, repetitive opinions of which the same user publishes are ignored. The topic growth can be evaluated by using the following equation:

$$Growth(x) = \left(\mu_1 * \sum_{i=1}^{n} f(m_i) + \mu_2 * \sum_{i=1}^{n} f(l_i) + \mu_3 * \sum_{i=1}^{n} f(c_i) + \mu_4 * \sum_{i=1}^{n} f(s_i)\right) * \frac{1}{n} \quad (12)$$

$$f(p_i) = \begin{cases} 1, & i = 1 \\ norm(\frac{p_i}{p_{i-1}+0.1}), & i > 1 \end{cases} \tag{13}$$

$$norm(y) = \begin{cases} 1, & y \geq 1 \\ y, & otherwise \end{cases} \tag{14}$$

$$\mu_1 + \mu_2 + \mu_3 + \mu_4 = 1 \tag{15}$$

where $Growth(x)$ is the growth degree of topic x, m_i is the number of the repliers of topic x within the ith time unit, l_i is the number of the opinion leaders of topic x within the ith time unit, c_i is the number of replies to topic x within the ith time unit, s_i is the number of opinions on topic x within the ith time unit, n is the total number of time units, μ_1 is the growth coefficient of a user, μ_2 is the growth coefficient of an opinion leader, μ_3 is the growth coefficient of a reply, and μ_4 is the growth coefficient of an opinion. The growth degree of a topic varies from 0.0 to 1.0, and the topic is at the maturity phase if the value is close to 1.

The topic hotness is evaluated by using the following equation:

$$Hotness(x) = \frac{cu}{n} * Growth(x) * (\lambda * sc(x) + \xi * qu(x)) * \Delta t(x)^{-k} \tag{16}$$

where $Hotness(x)$ denotes the hotness of topic x, n is the total number of time units, cu is the number of the consecutive time units in which topic x occurs, $Growth(x)$ is the growth degree of topic x, $sc(x)$ is the total number of replies to topic x, $qu(x)$ is the quotation number of the posts that belong to topic x, $\Delta t(x)$ is the time difference between the publication date of topic x and the current time, λ is the comment coefficient, ξ is the quotation coefficient of a post, and k is the decay coefficient.

5 Experiments and discussion

Experiments are performed in order to validate the feasibility and effectiveness of the proposed method. At the same time, the influence of opinion leaders on topic propagation as well as their features is analyzed. The test sample set includes the 1520 posts and 202290 related replies published at the China Sina blog website [37]. The publication dates of the test samples are between November 9, 2011 and January 18, 2012. The training sample set includes 17910 plain texts for the keyword extraction from the China Sogou laboratory [38], 14317 Web comments for the sentiment analysis, and the 12 blog hot topics that are listed at the China Sina blog website in 2011 and refer to the society, politics and economy. The software ICTCLAS is applied for Chinese word segmentation [20].

5.1 Performance of the proposed method

The named entity recognition is a big problem in natural language processing. The same problem exists in the arrangement of blog data too. In order to improve the precision of the named entity recognition of ICTCLAS, a user dictionary is constructed according to the tags of posts manually because such tags can be successfully used

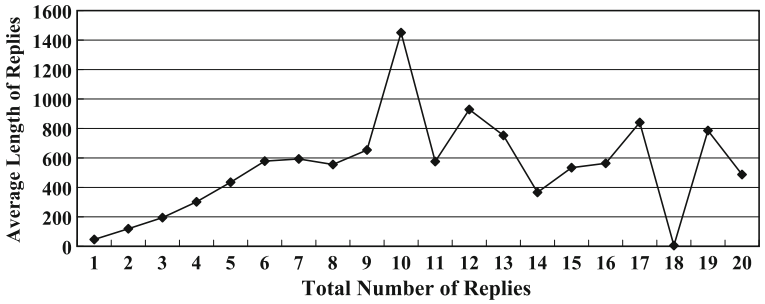

Figure 9 Reply characteristics of pseudonymous users

to detect the bursty event [46]. Moreover, anonymous users frequently participate in the opinion interaction and are more likely to publish negative opinions [21]. Hence, the characteristics of anonymous users need to be analyzed for knowing the structure of a blog community. It is proven that most of the bloggers are inclined to have a username without revealing the real name [21]. Hence, the characteristics of anonymous users can be inferred by observing the behavior patterns of the users that use pseudonymous username. The number of replies and the average length of replies for different kinds of pseudonymous users categorized by the total number of replies are counted. As shown in Figure 9, the pseudonymous users in the training sample set are more likely to continue commenting on the topic when the previous replies of the users are long. A comment threshold is consequently set to estimate the range of the anonymous users within the topic group, and its value is set as 200.

Experiment 1 evaluates the topic hotness on the basis of the proposed method, but not takes the degree of topic growth into account. Experiment 2 adopts the proposed method to evaluate the topic hotness. Experiment 3 applies the hot topic detection method based on the agglomerative hierarchical clustering algorithm [14], and the topic hotness is evaluated by counting the total number of posts and replies. In order to select optimum parameters for the performance of the proposed method, the influence of parameter regulation on the performance is observed at first. As shown in Figure 10, the comment coefficient λ is set as 0.1 and the quotation coefficient of post ξ is set as 0.5. As shown in Figure 11, the performance of the proposed method

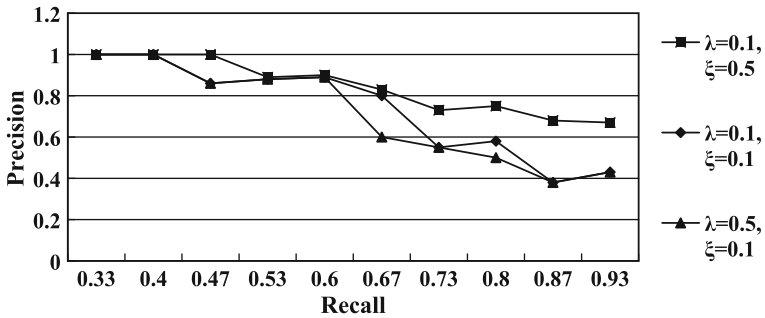

Figure 10 The performance comparison under different parameters for the proposed method

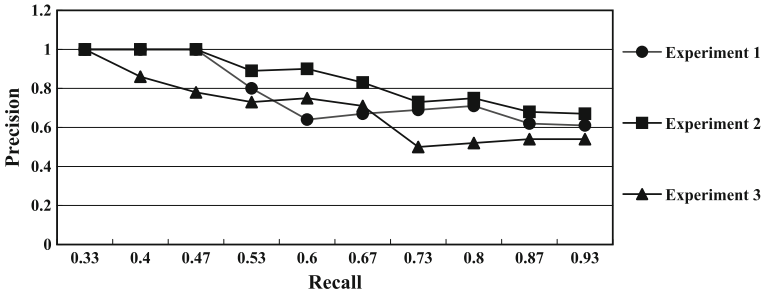

Figure 11 The performance comparison of three experiments

is good. The agglomerative hierarchical clustering algorithm is based on the vector space model, and most of the blog posts are not normalized so that the accuracy of the topic detection method based on the agglomerative hierarchical clustering algorithm is low. However, the topic detection based on the view of event reports focuses on keywords of a post. Hence, the posts that are not normalized do not cause a great impact on the proposed method. Moreover, the precision of hot topic detection is improved by measuring the growth state of a blog topic. On the other hand, the topics recommended at the blog website are all detected in three experiments. Hence, the influence of the blog service provider on topic propagation cannot be ignored.

The reply, replier, opinion leader and network opinion are four important factors for evaluating the development trend of a blog topic in the paper. The opinion leaders and network opinions are often ignored by traditional methods, and the influence of opinion leaders and network opinions on the performance of the proposed method needs to be validated. Hence, the following three conditions are considered. Condition 1 takes the above four factors into account when topics are evaluated by using the topic hotness evaluation equation listed in Section 4.4. Condition 2 ignores the influence of opinion leaders. Condition 3 ignores the influence of opinion leaders and network opinions. The performance comparison is as shown in Figure 12, the result of hotness evaluation is good when the four factors are taken into account. The topic drift often happens during the dispute so that many replies are not useful in evaluating the topic hotness. On the contrary, the network opinions on the event attributes or the event can exactly reflect the state of a topic. Moreover, the opinion

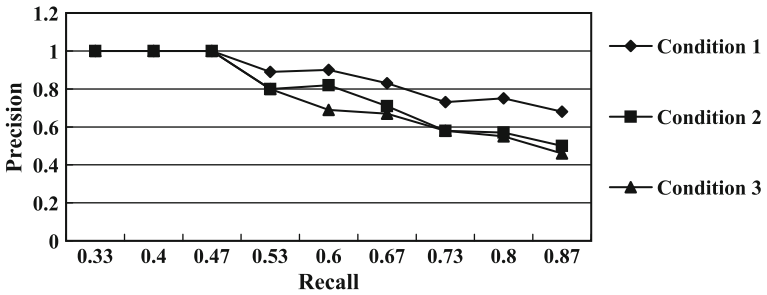

Figure 12 The performance comparison under three conditions

leaders play a vital role in the formation of the network opinions, and their opinions are representative. Hence, the performance can be better if the opinion leaders are taken into account too.

5.2 Discussion on opinion leaders

An opinion leader must actively publish high quality posts to maintain the status. Hence, it is very important to measure the quality of a post for recognizing the opinion leader. In order to test the improved method stated in Section 4.3.2 for ranking posts, the 378 posts which are related to 9 hot topics and include the 86 posts recommended by the Sina website are extracted from the training sample set. The method proposed by Agarwal et al. [1] is used as a baseline. The experimental result is shown in Figure 13. Although some bloggers close the function of post reply, the improved method considers the post quotation to counteract the effect of lack of comment features. Moreover, the links between posts are distinguished and filtered to avoid the topic drift. Hence, the improved method for post ranking has a better performance.

Famous bloggers in a given domain often have a number of followers and more advantages than ordinary users. In order to identify if the famous blogger is an opinion leader or not, an experiment is carried out by using the 1406 famous bloggers from 14 domains picked at the Sina blog website, and the opinion leaders who appear in the training sample set are identified by using the influence evaluation equation stated in Section 4.3.3. The behavior of opinion leaders and famous bloggers in different time intervals is observed. As shown in Figure 14, the social networks in different time intervals are constructed on the basis of the interaction relationships among opinion leaders and famous bloggers. Circle nodes in the social network stand for the opinion leaders within the current time interval. Rectangle nodes stand for the famous bloggers or previous opinion leaders, and an arrow points to a responder. Figure 14 represents the behavior patterns of opinion leaders and famous bloggers during the development of the topic related to the news that bin laden is killed. Those experimental results show that long-term opinion leaders are in the minority in comparison with other kinds of opinion leaders. The famous blogger in a given domain is not always an opinion leader. Opinion leaders prefer to influence other users by means of a blog post because they seldom take part in opinion interaction in other blogs. In order to analyze the influence of an opinion leader during the

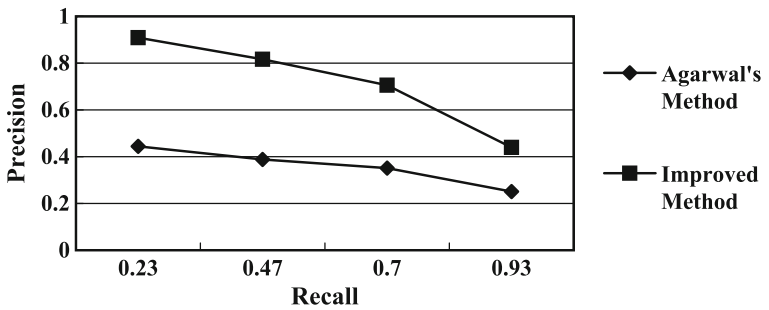

Figure 13 The performance comparison between two methods in post ranking

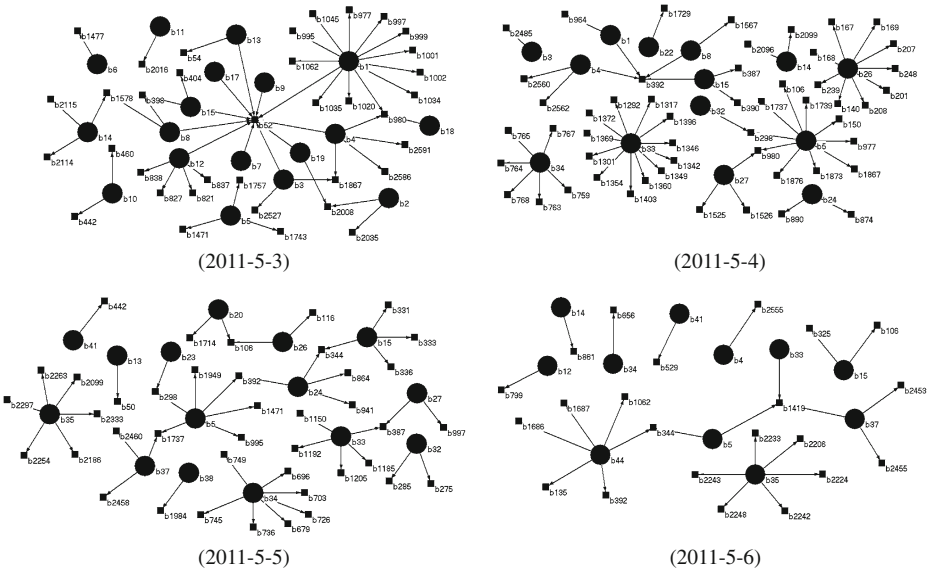

Figure 14 The interaction among opinion leaders and famous bloggers within different time intervals

topic propagation, topic propagation networks within different time intervals are constructed. The information propagator who publishes the origin report and plays the role of an opinion leader in the current or previous phases is labeled in the form of a circle node, and a rectangle node stands for an information receiver in the network. The quotation of a blog post is represented as an edge between the propagator and the receiver. Figure 15 shows the propagation process of the topic related to the news that bin laden is killed, and the post quotation of each propagator shows a downward trend. According to the experimental results, the influence of opinion leaders during the topic propagation gradually becomes weak. Hence, opinion leaders must capture the latest information and publish sharp remarks to keep their status.

The reasons for above phenomena can be explained as follows. The real identities of bloggers are ignored so that the credibility of information is based on the reasonable explanation and the development status of the related event. Hence, the formation of an opinion leader shows the rapid and unstable features. At the same time, with the improvement of retrieval engineers and the attention of a large number of Web media, it is easy for a user to know the comprehensive information. Moreover, the authorities often actively publish the latest investigation results on the hot event. The influence of opinion leaders on topic propagation is consequently weak.

The status of an opinion leader is formed during the information propagation. As for an opinion leader, it is an important problem to make sure other users receive and accept his/her information. Hence, what influences the user choice of information becomes more and more important. As for the work of Consutantiou and colleagues, different heuristics are adopted to understand the user choice for online news from the percentages of recommendations and readership, information source, the characteristics of a text and a picture [13]. Their study found that the percentages

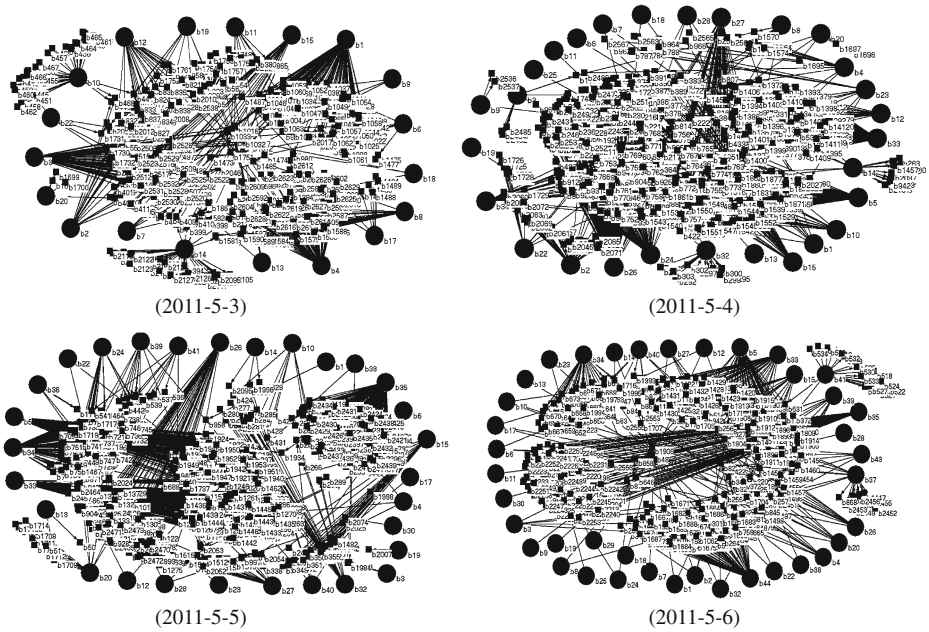

Figure 15 The propagation process of the hot topic related to the news that bin laden is killed

of recommendations and readership as well as the information resource's reputation have a positive effect on the user choice. However, our experimental results have some different findings. Our research observes the influence of opinion leaders on topic propagation by means of the information propagation network based on post quotation. Our experimental results show that the quotation behavior of bloggers mainly occurs within a very limited time, and the influence of a post gradually becomes weak with the lapse of time. On the other hand, bloggers seldom quote many related posts at the same time, and the most of bloggers prefer to choose the blog whose owner has the same interest and opinion. Hence, our research shows that the topic novelty has a great impact on the user choice and posts that own the high percentages of recommendations and readership usually do not influence bloggers for a long time. User interests also have a great impact on the user choice, which can explain why bloggers sometimes prefer the news blog to authoritative sources.

6 Conclusions and future work

Information overload is a big problem for Web mining. However, some representative features of hot topics can be reflected in some correlated blogs, and the propagation mechanism of a blog topic causes that users play a vital role in the formation and development of the topic. Hence, it is useful to analyze the phenomena that appear in those correlated blogs based on the W2T methodology for detecting hot topics. A news blog topic is closely related to a news event and the personal desire of users. In other words, the content of such a topic is composed of the narration of an event and

discussion on the event. The theme of a post reflects the personal desire or opinion of a blogger, and bloggers often publish the themes with respect to a news event from different views or levels. The blog information is consequently dealt with in different information granularities. The temporal feature is a typical feature of news blog topics so that the evaluation results of topic hotness are different in different life phases. Hence, it is important to identify what determines the development trend of a blog topic. The influence of an opinion leader is amplified by the Web platform, even impacts on the development of the related new event. The formation characteristics of opinion leaders consequently contribute to correctly analyzing their formation mechanism and effectively evaluating their influence. The major contributions of the paper are as follows:

- A topic detection approach based on the view of event reports is proposed, which takes into account the information granularity and evolutionary relationship between events to construct a news blog topic model.
- A method for identifying an opinion leader in blogspace is proposed, which carefully measures the popularity and recognition of a blog user.
- An evaluation measure for the growth state of a topic is proposed for detecting the current and forthcoming hot topic, which takes the changes of repliers, opinion leaders, replies and network opinions into consideration.

Experimental results show that the proposed method is feasible and effective. However, there are still some shortcomings as future work. First, experiments do not consider synonymous words. Second, the evolutionary relationship between events can be further decomposed into the temporal and causal relationships, and the causal relationship between events is not taken into account. Third, some posts are very popular because some impressive pictures or video are embedded into the posts. Hence, the influence of a picture or video on a user needs to be analyzed. At last, the human nature formed in the physical world still has a great impact on the virtual network, although the virtual network breaks some barriers from the physical world. The behavior patterns of different users in blogspace consequently need to be further studied in the future, and the theories of sociology or psychology can be adopted to understand the user behavior. At the same time, the user personality also needs to be well analyzed because the user behavior reflects his/her own need and interest. Moreover, the information is reprocessed and refined by users during the information propagation. If such information is extracted according to the human information processing mechanism, the problem solving strategy may be more effective and efficient. In the future, the shortage of word processing will be improved. Moreover, we will pay more attention to the latest investigation on the social contact patterns in blogspace.

Acknowledgements The study was supported by National Natural Science Foundation of China (60905027) and Beijing Natural Science Foundation (4102007).

References

1. Agarwal, N., Liu, H., Tang, L.: Identifying the influential bloggers in a community. In: Proceedings of the International Conference on Web Search and Web Data Mining, pp. 207–217 (2008)

2. Akritidis, L., Katsaros, D., Bozanis, P.: Identifying the productive and influential bloggers in a community. IEEE Trans. Syst. Man Cybern. **41**(5), 759–764 (2011)
3. Allan, J., Papka, R., Lavrenko, V.: On-line new event detection and tracking. In: Proceedings of the Twenty-First Annual International ACM SIGIR Conference, pp. 37–45 (1998)
4. Anderson, J.R., Schooler, L.J.: Reflections of the environment in memory. Psychol. Sci. **2**(6), 396–408 (1991)
5. Balakrishnan, H., Deo, N.: Discovering communities in complex networks. In: Proceedings of the Forty-Fourth Annual Southeast Regional Conference, pp. 280–285 (2006)
6. Bansal, N., Chiang, F., Koudas, N., Wm, F.: Seeking stable clusters in the blogosphere. In: Proceedings of the Thirty-Third International Conference on Very Large Data Bases, pp. 806–817 (2007)
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
8. Bodendorf, F., Kaiser, C.: Detecting opinion leaders and trends in online social networks. In: Proceedings of the Fourth International Conference on Digital Society, pp. 124–129 (2010)
9. Brants, T., Chen, F., Ioannis, T.: Topic-based document segmentation with probabilistic latent semantic analysis. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management, pp. 211–218 (2002)
10. Cao, Y.Z., Shao, P.J., Li, L.Q.. Topic propagation model based on diffusion threshold in blog networks. In: Proceedings of 2011 International Conference on Business Computing and Global Information, pp. 539–542 (2011)
11. Chen, C.C., Chen, Y.T., Chen, M.C.: An aging theory for event life-cycle modeling. IEEE Trans. Syst. Man Cybern. **37**(2), 237–248 (2007)
12. Chen, K.Y., Luesukprasert, L., Chou, S.C.T.: Hot topic extraction based on timeline analysis and multidimensional sentence modeling. IEEE Trans. Knowl. Data Eng. **19**(8), 1016–1025 (2007)
13. Constantiou, L., Hoebel, N., Zicari, R.V.: How do framing strategies influence the user's choice of content on the web. Concurrency Comput. Pract. Exper. **24**(17), 2207–2220 (2012)
14. Dai, X.Y., Chen, Q.C., Wang, X.L., Xu, J.: Online topic detection and tracking of financial news based on hierarchical clustering. In: Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, vol. 6, pp. 3341–3346 (2010)
15. Ding, F.: Research on information interaction and diffusion in internet communities. Beijing Jiaotong University, Beijing (2010)
16. Gong, H.J.: Research on automatic network hot topics detection. Central China Normal University, Wuhan (2008)
17. He, T.T., Qu, G.Z., Li, S.W., Tu, X.H., Zhong, Y., Ren, H.: Semi-automatic hot event detection. In: Proceedings of the Second International Conference on Advanced Data Mining and Applications, pp. 1008–1016 (2006)
18. Hong, Y., Zhang, Y., Fan, J.L., Liu, T., Li, S.: New event detection based on division comparison of subtopic. Chinese Journal of Computers **31**(4), 687–695 (2008)
19. Huang, H.H., Kuo, Y.H.: Cross-lingual document representation and semantic similarity measure a fuzzy set and rough set based approach. IEEE Trans. Fuzzy Syst. **18**(6), 1098–1111 (2010)
20. ICTCLAS. Home page: http://ictclas.org. Accessed 10 Mar 2011
21. Kilner, P.G., Hoadley, C.M.: Anonymity options and professional participation in an online community of practice. In: Proceedings of the 2005 Conference on Computer Support for Collaborative Learning, pp. 272–280 (2005)
22. Ku, L.W., Liang, Y.T., Chen, H.H.: Opinion extraction, summarization and tracking in news and blog corpora. In: Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, pp. 100–107 (2006)
23. Kumar, R., Novak, J., Raghavan, P.: On the bursty evolution of blogspace. World Wide Web **8**(2), 159–178 (2005)
24. Li, J.J., Zhang, X.C., Weng, Y., Hu, C.J.: Blog hotness evaluation model based on text opinion analysis. In: Proceedings of the Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, pp. 235–240 (2009)
25. Li, Y.M., Lai, C.Y., Chen, C.W.: Discovering influencers for marketing in the blogosphere. Inf. Sci. **181**(23), 5143–5157 (2011)
26. Lim, S.H., Kim, S.W., Park, S.J., Lee, J.H.: Determining content power users in a blog network: an approach and its applications. IEEE Trans. Syst. Man Cybern. **41**(5), 853–862 (2011)
27. Liu, Y., Yu, X.H., An, A.J., Huang, X.J.: Riding the tide of sentiment change: sentiment analysis with evolving online reviews. World Wide Web. doi:10.1007/s11280-012-0177-1

28. Luo, H.: A study on the evolution of internet public opinion of social focused events. Huazhong University of Science and Technoloy, Wuhan (2011)
29. Ma, X.H., Li, L.: Why do people blog? exploration of motivations for blogging. In: Proceedings of the Second IEEE Symposium on Web Society, pp. 119–122 (2010)
30. Musial, K., Budka, M., Juszczyszyn, K.: Creation and growth of online social network how do social networks evolve? World Wide Web. doi:10.1007/s11280-012-0179-z
31. Musial, K., Kazienko, P.: Social networks on the internet. World Wide Web **16**(1), 31–72 (2013)
32. Pan, X.: Opinion spreading models on complex network. Dalian University of Technology, Dalian (2010)
33. Qi, H.F.: Research on hot topic detection and event tracking in network public opinion. Harbin Engineering University, Harbin (2008)
34. Qiu, H.M.: The social network analysis of blogosphere. Harbin Institute of Technology, Harbin (2007)
35. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management **24**(5), 513–523 (1988)
36. Shi, J., Hu, M., Dai, G.Z.: Topic analysis of Chinese text based on small world model. Journal of Chinese Information Processing **21**(3), 69–75 (2007)
37. Sina Blog Website. Home page: http://blog.sina.com.cn. Accessed 1 Feb 2012
38. Sogou Laboratory. Home page: http://www.sogou.com/labs/dl/c.html. Accessed 28 Oct 2009
39. Song, X.D., Chi, Y., Hino, K., Tseng, B.: Identifying opinion leaders in the blogosphere. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, pp. 971–974 (2007)
40. Sun, W.J., Qiu, H.M.: A social network analysis on blogospheres. In: Proceedings of 2008 International Conference on Management Science and Engineering, pp. 1769–1773 (2008)
41. Wang, C.H., Zhang, M., Ma, S.P., Ru, L.Y.: Automatic online news issue construction in web environment. In: Proceedings of the Seventeenth International Conference on World Wide Web, pp. 457–466 (2008)
42. Wang, J.H.: Web-based verification on the representativeness of terms extracted from single short documents. In: Proceedings of 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, vol. 3, pp. 114–117 (2011)
43. Wang, Y., Xi, Y.H., Wang, L.: Mining the hottest topics on Chinese webpage based on the improved k-means partitioning. In: Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, pp. 255–260 (2009)
44. Xie, G.H.: The research on the system of the affect of internet opinion leaders. Central China Normal University, Wuhan (2011)
45. Yang, C.C., Shi, X.D., Wei, C.H.: Discovering event evolution graphs from news corpora. IEEE Trans. Syst. Man Cybern. **39**(4), 850–863 (2009)
46. Yao, J.J., Cui, B., Huang, Y.X.: Bursty event detection from collaborative tags. World Wide Web **15**(2), 171–195 (2012)
47. Yao, J.T., Yao, Y.Y.: Information granulation for web based information retrieval support systems. In: Proceedings of the Society of Photo-Optical Instrumentation Engineers, vol. 5098, pp. 138–146 (2003)
48. Yao, Y.Y., Petty, S.: Multiple representations of web content for effective knowledge utilization. In: Proceedings of 2012 International Conference on Brain Informatics, pp. 338–347 (2012)
49. Yu, H.: Research on the opinion leaders of political BBS: an case study on Sino-Japan BBS of strong nation forum. Huazhong University of Science and Technology, Wuhan (2007)
50. Zhang, Y.: A study on the phenomenon of public-opinion-spreading through bulletin board system. Jilin University, Changchun (2011)
51. Zhang, Y.C., Liu, Y., Ding, F., Si, X.M.: The research on stability of diffusion and competition between online topics. Int. J. Mod. Phys. C **21**(12), 1517–1529 (2010)
52. Zhao, J.: Web usage mining based on granularity computing. South China University of Technology, Guangzhou (2010)
53. Zhao, K., Kumar, A.: Who blogs what: understanding the publishing behavior of bloggers. World Wide Web. doi:10.1007/s11280-012-0167-3
54. Zhao, P., Cai, Q.S., Wang, Q.Y., Gen, H.T.: An automatic keyword extraction of Chinese document algorithm based on complex network features. Pattern Recognition and Artificial Intelligence **20**(6), 827–831 (2007)

55. Zhong, N., Bradshaw, J.M., Liu, J.M., Taylor, J.G.: Brain informatics. IEEE Intell. Syst. **26**(5), 16–21 (2011)
56. Zhong, N., Ma, J.H., Huang, R.H., Liu, J.M., Yao, Y.Y., Zhang, Y.X., Chen, J.H.: Research challenges and perspectives on Wisdom Web of Things (W2T). J. Supercomput. 1–21 (2010). doi:10.1007/s11227-010-0518-8
57. Zhou, Y.D., Sun, Q.D., Guan, X.H., Li, W., Tao, J.: Internet popular topics extraction of traffic content words correlation. Journal of Xian Jiaotong University **41**(10), 1142–1145 (2007)
58. Zhu, M.X., Cai, Z., Cai, Q.S.: Automatic keywords extraction of Chinese document using small world structure. In: Proceedings of Natural Language Processing and Knowledge Engineering, pp. 438–443 (2003)
59. Zhu, T.: Research on node role and group evolution in social network. Beijing University of Posts and Telecommunications, Beijing (2011)