Hot Topic Detection in News Blogs from the Perspective of W2T

Erzhong Zhou[1], Ning Zhong[1,2], Yuefeng Li[3], and Jia-jin Huang[1]

[1] International WIC Institute, Beijing University of Technology
Beijing 100124, P.R. China
{zez2008,hjj}@emails.bjut.edu.cn
[2] Department of Life Science and Informatics, Maebashi Institute of Technology
460-1 Kamisadori-Cho, Maebashi 371-0816, Japan
zhong@maebashi-it.ac.jp
[3] Faculty of Science and Technology, Queensland University of Technology
Brisbane QLD 4001, Australia
y2.li@qut.edu.au

Abstract. News blog hot topics are important for the information recommendation service and marketing. However, information overload and personalized management make the information arrangement more difficult. Moreover, what influences the formation and development of blog hot topics is seldom paid attention to. In order to correctly detect news blog hot topics, the paper first analyzes the development of topics in a new perspective based on W2T (Wisdom Web of Things) methodology. Namely, the characteristics of blog users, context of topic propagation and information granularity are unified to analyze the related problems. Some factors such as the user behavior pattern, network opinion and opinion leader are subsequently identified to be important for the development of topics. Then the topic model based on the view of event reports is constructed. At last, hot topics are identified by the duration, topic novelty, degree of topic growth and degree of user attention. The experimental results show that the proposed method is feasible and effective.

1 Introduction

A blog is an online diary which provides information sharing and opinion interaction service. A topic is defined as a seminal event or activity, along with all directly related events and activities [3]. A blog hot topic is a topic which users widely discuss and consecutively pay attention to in blogspace. Nowadays, most scholars pay more attention to overcoming the technique bottleneck because information overload and personalized management increase difficulties in topic detection.

Hokama et al. [6] adopted an extended agglomerative hierarchical clustering technique regarding the timestamp to detect topics, and evaluated the topic hotness by counting the number of articles related to each topic. He et al. [5] used

R. Huang et al. (Eds.): AMT 2012, LNCS 7669, pp. 22–31, 2012.

incremental term frequency inverse document frequency model and incremental clustering algorithm to detect new events, and considered the frequency and consecutive time of news reports to evaluate the topic hotness. Chen et al. [3] first extracted hot words based on the distribution over time and life cycle, then identified key sentences and grouped the key sentences into clusters that represent hot topics. As for evaluating blog topics, the methods mentioned above have many shortcomings. First, most of evaluation strategies focus on not users but media. Second, the hotness evaluation measures are too simple, and many factors which influence topics are not carefully analyzed. For example, a temporal social network is formed during the topic propagation and opinion interaction. Roles of blog users in topic propagation need analyzing because different kinds of users have different impacts on blog topics. Hence, what determines the formation and development trend of a blog topic is still an unsolved problem.

W2T (Wisdom Web of Things) proposed by Zhong et al. is an extension of the Wisdom Web in the Internet of Things age [13]. Besides, the W2T methodology gives a perspective on how to unify humans' models, information networks and granularity for analyzing the intersectional problems between the social world and cyber world [14]. In order to identify the important factors for detecting 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 different kinds of complex networks related to users. Then the influence of different kinds of users and network opinions on blog topics is measured in related information networks corresponding to the special information granularity. 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 analyzes some problems related to the news blog topics. Section 3 describes the proposed method. Section 4 tests the feasibility and effectiveness of the proposed method by some experiments. Section 5 gives conclusions and future work.

2 Problem Analysis

In order to construct an effective topic model to represent news blog topics and reasonably evaluate the topics, the paper has to cope with the two problems ignored by the traditional methods. First, the characteristics of news blog topics need to be analyzed on the basis of the motivations and features of blog users. Second, the context of topic propagation needs to be decomposed into the related complex networks, and the key factors that determine the development trends of news blog topics are identified in different complex networks.

2.1 Characteristics of a News Blog

The characteristics of information are influenced by the related medium. For example, there is no limitation on the writing style of a blog post, and users often

publish a topic from different views or levels. A blog topic model is consequently dynamic and hierarchical. Blogs can be categorized by the types of topics, and users visit different kinds of blogs with different motivations [8]. Hence, the development of a blog topic is related to the characteristics of the related blog. A news blog is a kind of temporal topic blog. A news blog topic is derived from the news event in the daily life, and often shows the burst and temporal features. Users take part in the 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 group is mainly maintained by user interests [11]. The topic group dissolves when users lose interests in the related topic.

2.2 Development of a Blog Topic

A blog topic can experience the birth, growth, maturity and fade [12]. 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. While the topic group rapidly grows up, some ordinary users can become opinion leaders that are influential in giving birth to the public opinion. When the topic grows up into the maturity, the public opinion emerges. At last, the influence of opinion leaders gradually becomes weak and then the topic fades away.

According to the description mentioned above, the user, topic and opinion are interrelated with each other in blogspace. Hence, the blog community, topic network and opinion network are extracted from the blogspace to analyze the context of topic propagation. The blog community is composed of users with the similar interest. The topic network is composed of topics with different granularities and the evolutionary relationship between topics. The opinion network is composed of opinions and interaction relationships among users.

As shown in Figure 1, the structure of the topic network can present compositions of each topic and the evolution of different topics. Besides, the boundary of a blog community is identified by blog topics. As a result, the information granularity not only determines the structure of the topic network but also identifies the range of the blog community. According to the social network analysis theory, the structure of an online community is related to user interaction patterns and relationships among users. The obvious phenomena are the most of users are active in the early phase. The phenomena can be explained as follows. Users lack the 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 comes into being, the phenomenon of 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. The structural changes of a blog community consequently reflect the development trends of topics. As far as the user role is concerned, some users become opinion leaders during the topic propagation and opinion interaction. Opinion leaders are influential during the evolution of network opinions. 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. Hence, the position of a blog user in the social network indicates the user role [10]. An opinion is a product of topic propagation. The formation and structural change of an opinion network all result from the interaction between users. The opinion expression is also the motivation of topic interaction, and the opinion network contributes to understanding the user behavior and evolution of network opinions. On the other hand, users also publish opinions on a topic from different views, and the opinion distribution in the opinion network can reflect the state of the topic. Besides, the structure and sentimental polarity distribution of the opinion network contribute to recognizing the opinion leaders [2]. Hence, the user behavior pattern, network opinion and opinion leader all impact on the development of a blog topic. Moreover, opinion leaders determine the development trend of the blog topic to a great extent.

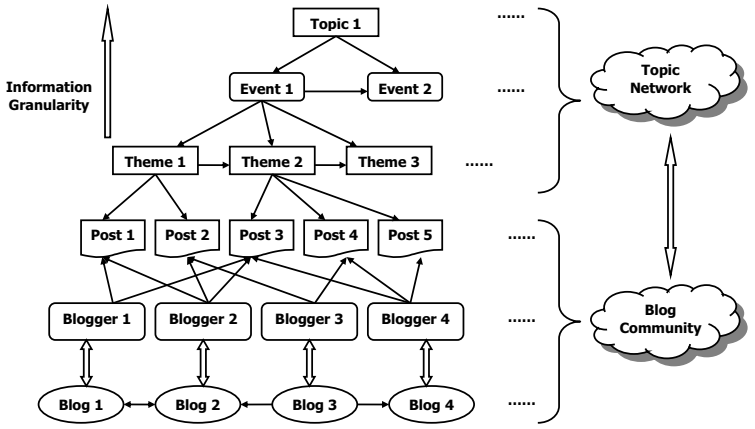

Fig. 1. The structure of a topic network in blogspace

3 Proposed Method

The information granularity needs considering for the construction of a topic model so as to analyze the structural characteristics and evolution of the topic. Hence, this section presents a new topic detection approach based on the different views of event reports and evolutionary relationship between events. As for the topic hotness, the changes of user interests can influence the vitality of the topic, and user interests need correctly evaluating. In order to detect the current and forthcoming hot topics, the growth state of a topic is measured, and opinion leaders are identified. Finally, hot topics are identified by the duration, topic novelty, degree of topic growth and degree of user attention.

3.1 Hot Topic Detection Algorithm

The hot topic detection algorithm comprises the topic detection and hotness evaluation, and is presented in Algorithm 1.

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, and construct the theme model of i;
3. **end for**
4. **for** time unit $j = 0$ to n **do**
5. Construct event models, and add the models into event list EL_j within j;
6. **if** j is equal to 1 **then**
7. Construct topic models within j manually;
8. **else**
9. Construct a new topic model, or update the previous topic model according to the related event model in EL_j;
10. **end if**
11. Analyze the structure of the blog community constructed by interaction relationships among users;
12. Detect network opinions on each topic according to the topic models;
13. Identify opinion leaders with respect to each topic;
14. Evaluate all topics within j by the duration, topic novelty, degree of topic growth and degree of user attention;
15. Select the topic whose hotness is more than d, and add new hot topics into T_{set};
16. **end for**
17. Return T_{set}.

3.2 Topic Detection Based on the View of Event Reports

Considering the information granularity in a topic network, a topic can be considered to be a cluster of related events, and an event can be considered to be a cluster 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 theme model is defined as $Theme = \{Keyword_1, Keyword_2, \ldots, Keyword_s\}$, where $Keyword_i$ denotes the ith keyword;
- An event model is defined as $Event = \{Theme_1, Theme_2, \ldots, Theme_m\}$, where $Theme_i$ denotes the ith theme;
- A topic model is defined as $Topic = \{Event_1, Event_2, \ldots, Event_n\}$, where $Event_i$ denotes the ith event.

The event model is the key to constructing the topic model. The single pass clustering algorithm is widely used in event detection and tracking [1], which can

be explained as follows. A report is merged into an event if the content similarity between two objects is above a threshold. Otherwise, the report is considered to be the first report of a new event. Topic detection based on the view of event reports first adopts single pass clustering algorithm to extract events which occur in a given time interval. Then evolutionary relationship between events is identified by the content similarity between events and event distribution in different topics. At last, the topic model is created or updated by detecting the new event and tracking the previous event.

A keyword is a basic element for a theme model. The keyword is identified according to the weight of the word. The weight of a word is evaluated by the following equation in the Term Frequency Inverse Document Frequency (TFIDF) method [9]:

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

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 number of blog posts, and N_k is the number of blog posts where t_k appears.

A theme model is the basis of an event model. When two events are compared, each theme in one event model is compared with that in the other event model. Then the average of similarities between themes of events is regarded as a proof of the similarity evaluation of two events. However, theme models abstracted from different posts can own different event attributions. Sometimes, one theme model can be a subset of the other theme model in terms of event attributes. Hence, the intersection between theme models is essential to compare the similarity between theme models. On the other hand, the size of a theme model has a great impact on the intersection between theme models. When the size of a theme model is small, the intersection between theme models is also small even if one theme model is the subset of the other one. As for that phenomenon, the minimum of two theme models is also taken into consideration by using the following equation, in order to correctly compare the similarity between theme models:

$$sim(d_i, d_j) = \alpha * \frac{|d_i \cap d_j|}{|d_i \cup d_j|} + \beta * \frac{|d_i \cap d_j|}{min(|d_i|, |d_j|)} \tag{2}$$

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 between d_i and d_j, $d_i \cup d_j$ is the union of 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.

In general, one event may evolve into other events. Moreover, main attributes of events often change. However, time is the key to the identification of the evolutionary relationship between events. If events occur very closely and the content similarity between events is high, the events are likely to be correlated with each other. The evolutionary relationship between two events is identified by the following norm. If the most of similar events which occur before the target event belong to the same topic, the target event evolves from the related events.

3.3 Topic Hotness Evaluation Based on User Interests

The growth state indicates the development trend of the topic. Replies, repliers, opinion leaders and network opinions all have great changes during the development of the blog topic. Hence, the growth degree of a topic is measured by using the following equation:

$$Growth(x) = (\mu_1 * \sum_{i=1}^{n} f(m_i) + \mu_2 * \sum_{i=1}^{n} f(l_i) + \mu_3 * \sum_{i=1}^{n} f(c_i) + \mu_4 * \sum_{i=1}^{n} f(s_i)) * \frac{1}{n} \quad (3)$$

$$f(p_i) = \begin{cases} 1, & i = 1 \\ norm(\frac{p_i}{p_{i-1}+0.1}), & i > 1 \end{cases} \quad (4)$$

$$norm(y) = \begin{cases} 1, & y \geq 1 \\ y, & otherwise \end{cases} \quad (5)$$

where $Growth(x)$ is the growth degree of topic x, m_i is the number of repliers of topic x within the ith time unit, l_i is the number of opinion leaders of topic x within the ith time unit, c_i is the number of replies of 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. When opinions are counted, repetitive opinions of which the same user publishes are ignored.

The degree of user participation, position in the social network and influence on neighbors' opinions are assessed to recognize the opinion leader. Besides, the user role during the topic propagation and evolution of network opinions need analyzing. The influence of a user is evaluated by the following equation:

$$Opind_n(w,x) = \sum_{i=1}^{n} \frac{nop_i(w,x)}{Tnop_i(x)} * (\varphi * \frac{nos_i(w,x)}{Tnos_i(x)} + \psi * cen_i(w) + \delta * \frac{bp_i(w,x)}{Tp_i(x)}) \quad (6)$$

where $Opind_n(w,x)$ is the influence of user w on topic x within the nth time unit, i denotes the ith time unit, $nop_i(w,x)$ is the number of opinions that user w publishes on topic x within the ith time unit, $Tnop_i(x)$ is the total number of opinions that all users publish on topic x within the ith time unit, $cen_i(w)$ is the central degree of user w in the social network within the ith time unit, $nos_i(w,x)$ is the number of users who have the same opinion as user w within the ith time unit, $Tnos_i(x)$ is the total number of users who publish opinions on topic x within the ith time unit, $bp_i(w,x)$ is the number of posts that user w publishes with respect to topic x and are recommended by the website within the ith time unit, $Tp_i(x)$ is the total number of posts that all users publish with respect to topic x within the ith time unit, φ is the emotion coefficient, ψ is the position coefficient, and δ is the quotation coefficient.

The topic hotness reflects user interests on a topic. If a topic is very new, the user is more likely to be interested in the topic. If a topic spreads a long time, the topic can have a great probability to attract users. The representations of user interests can also be reflected from the quotation number and reply number of posts. Hence, the topic hotness is evaluated by using the following equation:

$$Hotness(x) = \frac{cu}{n} * Growth(x) * (\lambda * sc(x) + \xi * qu(x)) * \Delta t(x)^{-k} \quad (7)$$

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

4 Experiments and Discussion

Experiments are performed in order to validate the feasibility and effectiveness of the proposed method. The test sample set includes 1520 posts and 202290 related replies published in the China Sina blog website. The publication date of the test sample is between November 9, 2011 and January 18, 2012. The training sample set includes 12 blog hot topics listed by the China Sina blog website in 2011 and 17910 plain texts for the keyword extraction from the China sogou laboratory. The Chinese word segmentation software ICTCLAS is applied.

As for the user behavior, anonymous users frequently participate in the opinion interaction and are likely to publish negative opinions [7]. Hence, the characteristics of anonymous users need analyzing for knowing the structure of a blog community. According to the related study, most of bloggers are inclined to have a pseudonymous username [7]. As a result, the characteristics of anonymous users can be inferred by observing the behavior patterns of the pseudonymous users. The number of replies and average length of replies for different kinds of pseudonymous users categorized by the total number of replies are counted. As shown in Figure 2, the pseudonymous user is more likely to continue replying when the length of the user reply is long. A reply threshold is consequently set to estimate the range of anonymous users. Experiment 1 evaluates the topic hotness on the basis of the proposed method but does not take the growth degree of a topic 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 [4], and topics are evaluated by the total number of posts and replies. As shown in Figure 3, the performance of the proposed method is good. The agglomerative hierarchical clustering algorithm is based on the vector space model, and most of blog posts are not normalized so that the accuracy of the topic detection method based on agglomerative hierarchical clustering algorithm is low. However, the proposed method

focuses on keywords of a post. Hence, blog 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.

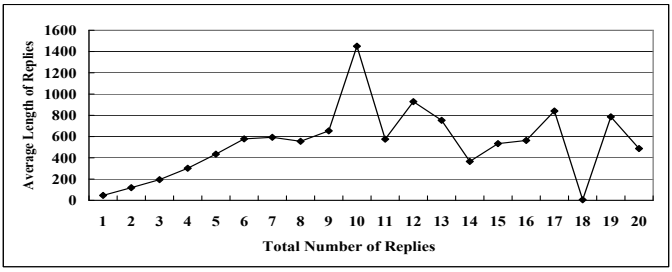

Fig. 2. Reply characteristics of pseudonymous users

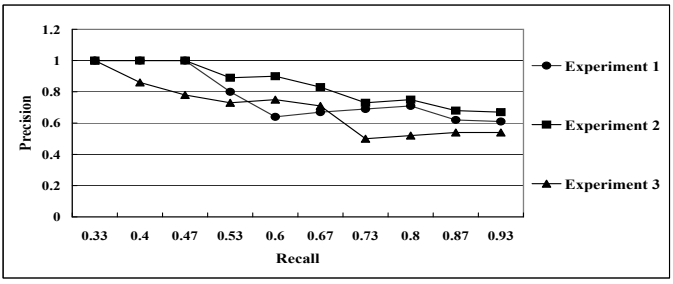

Fig. 3. Performance comparison of three experiments

5 Conclusions and Future Work

The problem solving strategy of W2T lays emphasis on human nature. The mechanism of information propagation in news blogs causes that the user motivation and behavior pattern determine the burst and temporal features of topics. Hence, users cannot be separated from topics in blog topic detection. Experimental results prove the proposed method is feasible and effective. The user behavior pattern, network opinion and opinion leader are identified to be important for the development trend of a blog topic. However, there are still some shortcomings. Experiments do not consider synonymous words. Although the virtual network breaks some barriers of the physical world, the human nature formed in the physical world still has a great impact on the virtual network. The behavior model of a blog user needs the further research. The shortage of word processing is going to be improved in the future. We will pay more attention to the latest researches on the model of the social contact in blogspace.

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

References

1. 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)
2. 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)
3. Chen, K.Y., Luesukprasert, L., Chou, S.C.T.: Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling. IEEE Transactions on Knowledge and Data Engineering 19(8), 1016–1025 (2007)
4. 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)
5. He, T., Qu, G., Li, S., Tu, X., Zhang, Y., Ren, H.: Semi-automatic Hot Event Detection. In: Li, X., Zaïane, O.R., Li, Z. (eds.) ADMA 2006. LNCS (LNAI), vol. 4093, pp. 1008–1016. Springer, Heidelberg (2006)
6. Hokama, T., Kitagawa, H.: Detecting Hot Topics about a Person from Blogspace. In: Proceedings of the Sixteenth European-Japanese Conference on Information Modeling and Knowledge Bases, pp. 290–294 (2006)
7. 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)
8. Qiu, H.M.: The Social Network Analysis of Blogosphere. Harbin Institute of Technology, Harbin (2007)
9. Salton, G., Buckley, C.: Term-weighting Approaches in Automatic Text Retrieval. Information Processing & Management 24(5), 513–523 (1988)
10. 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)
11. 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)
12. Zhang, Y.: A Study on the Phenomenon of Public-opinion-spreading Through Bulletin Board System. Jilin University, Changchun (2011)
13. 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). Journal of Supercomputing (2010)
14. Zhong, N., Bradshaw, J.M., Liu, J.M., Taylor, J.G.: Brain Informatics. IEEE Intelligent Systems 26(5), 16–21 (2011)