A Fast Method to Detect Hot Topic from BBS

XU Hui-jie CAI Wan-dong CHEN Gui-rong

School of Computer Science and Technology, Northwestern Polytechnical University, Xi'an 710129, China

Abstract

Due to the information of Bulletin Board System (BBS) is numerous, detecting hot topic from BBS is a challenging task. In this paper, we propose a novel method for fast detection of hot topic. By analyzing features of hot topic, we define the hotness indicator to filter out the non-hot threads. Then we fit the changeful curves of threads followed time and remove the threads whose vitality is on the decline. Finally, we give an algorithm that can merge the remaining threads and get hot topics. Results prove that the proposed method is of good validity and feasibility.

Keywords: BBS, hot topic, thread, aging theory

1. Introduction

Bulletin Board System (BBS), which is a public discourse space, is open to the public and no identity restriction. When some hot topics appear in BBS, they often caused heated debate among BBS users. Sometimes, these topics may trigger a major public opinion crisis. Therefore, finding a method which can quickly and efficiently detect hot topic from BBS is of important significance to identify and monitor the network public opinion.

The traditional hot topic detection mainly uses semantic analysis and social network analysis, many later approaches are extension of the two methods. On the semantic analysis side, Ye et al. [1] and Li et al. [2] considered the number of posts or keywords as a reference to judge whether or not users give attention to a topic and established a candidate set of hot topics. Then, the weights of the candidate topics were calculated by the classical TF-IDF scheme. By carrying out multiple filtrations for the candidate set, they eventually achieved the hot topic detection in the BBS or blog environment. Due to the vast majority of TDT (Topic Detection and Tracking) solutions proposed for topic detection are not suitable for posts in web forums, Chen et al. [3] proposed a noise-filtered model to extract bursty topics from web forums using terms and participations of users. Xi et al. [4] proposed a topic tracking method for BBS data based on semantic similarity. This method firstly constructed keywords tables of topic and post as their representation models, and then computed the two tables' semantic similarity which is served as correlation degree between post and topic. Finally, this method used the correlation degree to realize BBS-oriented topic tracking. Further work includes hot topic detection for BBS based on the aging theory [5] [6] [7]. These methods first extracted candidate topics from BBS data by the clustering method. And then, according to the aging theory, the hotness of topics were valuated. Finally, topics were ranked and hot topics were detected. Along with the development of research, many researchers have turned their eyes on hot topic detection based on the complex network theory in recent years. Kleinberg [8] and Wang et al. [9] first constructed a user network based on user interest and then detected hot topic associated with those users. Zhuang [10] constructed a bipartite graph by considering users and posts as vertices and detected the post with the most important influence.

To some extent the method based on the semantic analysis and complex network analysis can be employed to detect hot topic within a certain period of time, but these methods exist the following issues: (1) hot topic detection based on the semantic analysis mainly use clustering method. Confronted by the huge amounts of BBS data, the problem of the method have high computing cost. (2) hot topic detection based on the complex network extract hot topic from BBS by analyzing features of network. but the existing methods mainly take quantity accumulation of the characteristic parameters as a judgmental indicator while neglect the time change characteristic of network.

In order to solve the above two problems, a method, which can quickly detect hot topic from BBS, is

proposed. After thoroughly researching the structural characteristics of BBS and the characteristics of topic, the method adopts the stepwise refinement strategy, through carrying out noise filtration, word segmentation and thread merging, achieve hot topic detection. Experiments prove that the method is practical and effective.

The remainder of the paper is organized as follows. In section 2, we introduce BBS structure and topic characteristics. Section 3 presents how to detect hot topic quickly from BBS. In section 4, we perform an experiment and analysis based on the proposed method. Finally, we conclude our work in section 5.

2. BBS structure and topic characteristics

BBS, which is a web application, provides users with services such as posting specific or general text, commention and so on. BBS is composed of many boards, and BBS texts are organized in threads. A discussion between BBS users is first sponsored by one of user submit a post to BBS, the first post is called the entry post. If other BBS users are interested in the subject that the entry post talks about, they may reply to either the entry post or a previous reply post (refer to Fig.1). All these posts from the same discussion form a thread, a BBS topic is a set for all those threads from the same content. Topic level view is shown in Fig.2 [11]:

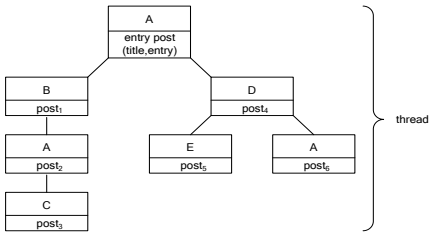

Fig.1: the Thread Structure in BBS

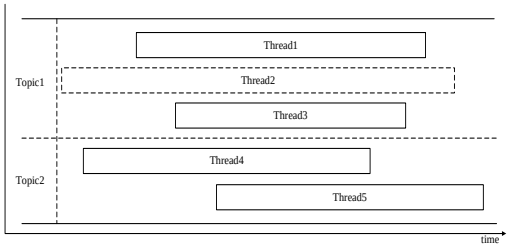

Fig.2: Topic Level View

3. Hot Topic Detection

In this paper, the proposed hot topic detection method includes four steps: noise filtration, word segmentation, topic combination and hot topic detection. Noise filtration is to carry out denoising for the existing BBS data. It is found out that most of the threads are general in BBS, while only a small proportion of threads can trigger extensive attention and fiery debate. Besides, according to the Aging theory, each topic experiences four different stages: birth, growth, decay and death [12], a thread in the decay stage will not evolve into a hot topic. Thus, by effectively filtering out the threads with the lower hotness and the threads which will not evolve hot topics, the amount of information that will be processed can be reduced to a minimum, the efficiency of information processing and recognition accuracy are therefore improved. After noise filtering, most of the threads, which will not evolve into hot topics, are filtered out. For the remaining threads, we use ICTCLAS (Institute of Computing Technology, Chinese Lexical the Analysis System) to extract terms from them. By running the clustering algorithm proposed in this paper, we may combine two or more threads into one topic—that is the hot topic. The rapid hot topic detection for BBS is illustrated in Figure 3.

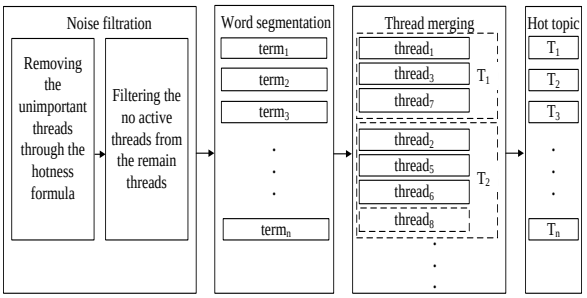

Fig.3: Rapid Hot Topic Detection for BBS

3.1. Noise Filtration

In this step, we tend to filter out the threads with the lower hotness and the threads which will not evolve into hot topics by indicator of hotness and hotness trend prediction so that hot topics can be detected quickly. Due to the title of the thread, namely the title of the entry post generally represents the theme of the thread [13], the remarkable propagation characteristic enlightens us, views of the entry post, replies and the resultant time series may be employed to find hot threads and predict the development trend of the hot threads. In this paper, we adopt the idea proposed by Lu [14] to score the hotness of the threads The calculation

formula of the hotness can be written as follows:

$$hotness(x_i) = \alpha \frac{r(x_i)}{avgr(X)} + \beta \frac{b(x_i)}{avgb(X)} + \gamma \frac{r(x_i)/b(x_i)}{\max(X)} \quad (1)$$

Where α, β, γ are the weight, $r(x_i)$ represents the replies of the entry post x_i, $b(x_i)$ represents the views of the entry post x_i, $avgr(X)$ represents the average replies of all posts X, $avgb(X)$ represents the average views of all posts X, $\max(X)$ represents the maximum ratio of the replies to views. In general, BBS users must first click a post before he/she reply to the post. Therefore, for a given post, its views are usually greater than its replies. In the equation, $r(x_i)/b(x_i)$ is used as an index of the intensity of the debate, The greater ratio of $r(x_i)$ to $b(x_i)$, the higher degree of concern that the post by the corresponding topic.

Before determining on above three indicators with respect to the degree of importance of each other, α, β and γ are unknown. In this paper, we adopt AHP method to solve the problem[15], based on Saaty "l-9" method, $\alpha = 0.3333$, $\beta = 0.1111$ and $\gamma = 0.5556$ can be obtained.

Above hotness score can quickly find top posts，but one can not be neglected that a thread in the decay stage will not evolve into a hot topic，Equation (1) reflects the replies of the post and views on the number of cumulative, but it does not reflect the hotness of post changes over time.

For instance, the discussion cycle of a post is very long, views and replies are also relatively great, the content what they discuss is not a current hot topic, but a problem which BBS users always concern over a long period of time. The change of the thread hotness should be a complex time series with the nonlinear rise or fall characteristic [16]. To avoid that, in this paper, we divide this time period from each entry post's birth to last reply into a series of time window. For each thread, we statistic their replies in all time window and carry out curve fitting to find those threads in the growth stage. For details, see the section 4.2.

By quantifying the indicator of hotness and predicting the development trend of hot post, we will filter out those threads with lower hotness and the threads which will not evolve into hot topics and remain the threads with high quality. And also, we will reduce the amount of BBS information at hand to the minimum. After noise filtration, we put the rest of the threads to establish a set $ST = \{th_1, th_2, ..., th_m\}$.

3.2. Word segmentation

Since several threads may contain same topic in real life. So before clustering the threads, we need to perform word segmentation for the titles of threads to extract nouns and verbs which reflect the characteristics of hot topics. Unlike English, Chinese content is consecutive, needs special Chinese word segmentation tool for processing. This paper uses ICTCLAS (Institute of Computing Technology，Chinese Lexical Analysis System) to carry out segmentation. For example, the title of the thread is like "朝鲜今日宣布正计划进行第三次核试验". By segmentation, we can get the segmentation results as follows: 朝鲜/n 今日/t 宣布/v 正/d 计划/v 进行/v 第三/m 次/q 核试验/n. After getting rid of repeated words and extracting nouns and verbs from it, we can obtain a set {朝鲜，宣布，计划，进行，核试验}.

In practice, for $ST = \{th_1, th_2, ..., th_m\}$, we perform word segmentation for the title of each element and establish a set $th_m = \{term_1, term_2, ..., term_n\}$, $term_i \neq term_j$ ($i \neq j$). Moreover, considering many topics in threads are new events in real life, such as "house sister" and "wristwatch brother". So before performing word segmentation, these new words need to be added manually into ICTCLAS dictionary.

3.3. Thread Merging

For th_i of $ST = \{th_1, th_2, ..., th_m\}$, we establish a set $th_i = \{term_j \mid 1 \le j \le n\}$. Since several threads may contain the same topic in real life, so th_i on the same topic will contain the same $term_j$. For instance, for three given sets th_1, th_2 and th_3, we may get the results by word segmentation as follows: $th_1 = \{term_1, term_2, term_3, term_4\}$, $th_2 = \{term_2, term_3\}$, $th_1 = \{term_1, term_2, term_3\}$. We define the Jaccard coefficient between the two sets as their similarity：

$$sim(th_i, th_j) = \frac{|th_i \cap th_j|}{|th_i \cup th_j|} \quad (2)$$

Where $th_i, th_j \in ST$.

Thread merging algorithm is as follows：

Input: $ST = \{th_1, th_2, ..., th_m\}$, threshold η.

Output: The set of the hot topic $H = \{H(th_k)\}$.

Step1: The set of the topic ST is initialized. (ST includes m threads).

Step2: For $th_1, th_2, ..., th_m$, $\max|th_i|$ is selected. If $\max|th_i|$ and th_j satisfy $sim(\max|th_i|, th_j) \geq \eta$, then $\max|th_i|$ and th_j are added into set $H(th_k)$, namely, $H(th_i) = \{th_i, th_j\}$, Otherwise, go to Step3.

Step3: $ST = ST - H(th_i)$, repeats steps (2). If $ST = \phi$, then the end of the computing.

4. Experiment Results and Analysis

4.1. Data set Description

This paper takes the WangYi News BBS (http://bbs.news.163.com) as the research object. With the help of the data acquisition program，we get the posting data between January 1, 2011 and April 30, 2011 and establish thread library. There are 6882 entry posts, 2346 entry post authors. The average replies of entry post is 8 times, the average views of entry post is 1755 times, the maximal replies of entry post is 745 times, the minimal views of entry post is one times.

4.2. Result Analysis

According to the above experimental methods, we score the hotness of all the entry posts by formula 1 and select top-20 posts. The results are shown in Table 1.

Table .1: Hotness Degree values of Top-20 Threads

Order	Thread	Hotness
1	湖南省宜章县赤石乡长城岭、塘背村民的哭诉(有图有真相)	32.51
2	中石油给日本捐款 3000 万	27.27
3	闵行区纪监委姑息养奸指鹿为马 竟把"捞模"当"劳模" 天理难容!	25.32
4	致上海市公安局局长张学兵的一封人民来信	22.16
5	为药家鑫的判决赌一把	18.46
6	原创长篇《画家古董商与情人》更名《戒.界》向茅盾文学奖进军	18.37
7	鸿门宴上的十三大谜团	14.55
8	满清统治者和日本侵略者的区别	14.06
9	揭秘刘志军落马的原因	13.81
10	鸡蛋修补耳膜引发的离奇官司——潜规则又露狰狞面孔	12.85
11	蒙古，满清为何得到中原江山	12.76
12	你做法官，如何判药家鑫？	12.70
13	拒绝死刑,挽救家鑫,现征集万民签名	11.00
14	对待日本地震的情感	9.66
15	新浪微博无辜封杀一剑传媒草根团队微博是可忍孰不可忍！	9.60
16	天价过路费法官调离，谁在恼羞成怒？	8.57
17	【时评】刘志军被查：人民公敌就应该是这样的下场！	8.36
18	葯家鑫案。孔慶東：這樣的事情，美國、日本絕不會有！	6.71
19	欢呼吧！药被判死刑！	6.17
20	日本地震，我看了看中国人的感想	6.00

It should be noted that the hotness score of thread is an accumulated value in 130 days. Therefore, in order to discover the development trend of threads, we take every 2 days as a time window and statistic the replies of threads in each time window. For the obtained data, we conduct curve fitting based on Gauss model (refer to Fig.4). Method proposed by Lu [14] is employed to judge the activity of thread in the future. The analysis reveals that the threads numbered 1、2、3、5、12、13、14、18、19、20 are in the active stage. Based on the proposed algorithm in section 2.2, We merge the above threads, the results are shown in table 2.

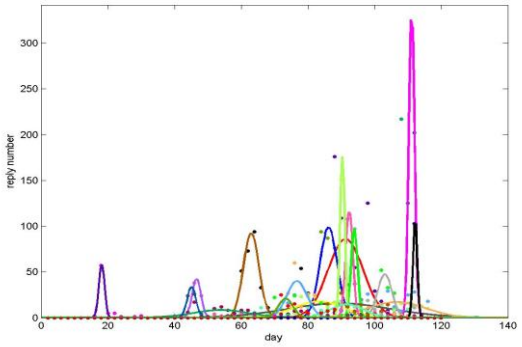

Fig.4: the development curves of the threads

Table.2: Hot topics from 2011-01-01 to 2011-04-30

Order	Terms	Hot topic
1	湖南省、宜章县、赤石乡、长城岭、塘北、村民、哭诉、有、图、真相	湖南省宜章县赤石乡长城岭塘背村民哭诉
2	闵行区、纪监委、姑息养奸、指鹿为马、把、捞摸、当、劳模、天理难容	闵行区纪监委把“捞模”当“劳模”
3	中石油、给、日本、捐款、地震、看、中国人、感想、对待、情感	中国给日本地震捐款
4	药家鑫、判决、赌、做、法官、判、拒绝、死刑、挽救、征集、万民、签名、案、孔庆东、事情、美国、日本、有、欢呼、死刑	药家鑫被判死刑

5. Conclusion

Because of the existing hot topic detecting methods are not suitable for BBS. In this paper, we propose a novel method for quickly detecting hot topic from BBS that appear in a specific time period. Our work makes two novel and important contributions:

- Compared with the traditional hot topic detection methods, we adopt a hot topic detection method based on progressive refinement. Before running the specific hot topic detection algorithm for BBS data, noise filtration is carried out to remove a lot of useless data.
- In order to improve the detection accuracy and reflect the dynamic characteristics of the popularity of hot topics, the ageing theory is introduced to our detection method.

The experiments demonstrate that our method improves the performance of hot topic detection in BBS environment on both efficiency and accuracy significantly.

6. References

[1] H. Ye, W. Cheng, G. Dai, "Design and Implementation of Online Hot Topic Discovery Model", *Wuhan University Journal of Natural Sciences*, vol. 11, no. 1, pp. 21–26, 2006.

[2] H. Li, H. Zhang and et al, "Keywords Based Hot Topic Detection on Internet", *The 5th National Information Retrieval Conference (CCIR 2009)*, pp.134-143, 2009.

[3] Y. Chen, S. Yang, X. Cheng, "Bursty Topics Extraction for Web Forums", *proceeding of the eleventh international workshop on Web information and data management*, pp. 55-58, 2009.

[4] Y. Xi, C. Lin and et al, "Method for BBS Topic Tracking Based on Semantic Similarity", *Journal of Computer Applications*, vol. 31, no. 1, pp. 93-97, 2011.

[5] D. Zheng, F. Li, "Hot Topic Detection on BBS Using Aging Theory", *Web Information System and Mining Lecture Notes in Computer Science*, pp. 129-138, 2009.

[6] H. Ma, "Hot topic extraction using time window", *Machine Learning and Cybernetics (ICMLC 2011)*, pp. 56-60, 2011.

[7] K.-Y. Chen, L. Luesukprasert, S.-C. T. Chou, "Hot Topic Extraction Based on Timeline Analysis and Multi-dimensional Sentence Modeling," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 8, pp.1016–1025, 2007.

[8] J. Kleinberg, "Bursty and Hierarchical Structure in Streams," *Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '02)*, pp. 91-101, 2002.

[9] L. Wang, G. Dai, "Forum Hot Topic Detection Based on Community Structure of Complex Networks", *Computer Engineering*, vol. 34, no. 11, pp. 214-216, 2008.

[10] M. Zhou, J. Zhuang, "The Application of Social Network Analysis in Management of BBS", *Technology Economics*, vol. 28, no. 11, pp. 93-98, 2009.

[11] M. Zhu, W. Hu and O. Wu, "Topic detection and tracking for threadeddiscussion communities", *In IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 77–83, 2008.

[12] C.C. Chen, Y.T. Chen, Y. Sun, and M.C. Chen, "Life Cycle Modeling of News Events Using Aging Theory," *Proc. 14th European Conf. Machine Learning (ECML '03)*, pp. 47-59, 2003.

[13] M. Lu, X. Yal, S. Wei, "BBS hot topic mining algorithm based on fuzzy clustering", *Journal of Dalian Maritime University*, vol. 34, no. 4, pp. 52-58, 2008

[14] J. Lu, H. Zhang, Y. Zhang, "Research on the Technology of Hot Topics Foundation and Trend Forecast in BBS", *Intelligent Computer and Applications*, vol. 2, no. 2, pp. 1-5, 2012.

[15] SAATY, T.L, "A scaling method for priorities in hierarchical structures," *Journal of Mathematical Psychology*, Vol. 15, No. 3, pp. 234-281, 1977.

[16] J. V. Hansen and R. D. Nelson, "Neural networks and traditional time series methods: A synergistic combination in state economic forecasts", *IEEE Trans. Neural Networks*, vol. 8, no. 4, pp. 863-873, 1997.