Hot Topic Detection on BBS Using Aging Theory*

Donghui Zheng and Fang Li

UDS-SJTU Joint Research Lab for Language Technology
Dept. of Computer Science and Engineering
Shanghai Jiaotong University, Shanghai, China
zhdhui@gmail.com, fli@sjtu.edu.cn

Abstract. BBS(Bulletin Board Systems) is one of the most common places for threaded discussion. It becomes more and more popular among web users, especially in China. Everyday a huge amount of new discussions are generated on BBS. It is too difficult to find hot topics. To solve this issue, we propose a novel approach to detect hot topics on BBS for any period of time. Our solution consists of three steps. First of all, candidate topics are extracted using the clustering method. Secondly, based on the extracted topics, aging theory is employed to valuate the hotness of topics. Both two steps above are carried out incrementally over time. Finally, topics are ranked and hot topics are detected. Experiments performed on practical BBS data show that our method is quite effective.

Keywords: Hot topic detection, Aging theory, BBS.

1 Introduction

As ReadWriteWeb[1] indicates, there are nearly 2 hundred million of BBS users in early 2008 in China. The total number of daily page views across BBS has reached over 1.6 billion, with 10 million posts published every day. However, too many topics are discussed on BBS. How to identify hot topics on BBS becomes more and more important.

Here are some basic concepts for BBS:

Post: A post is a user submitted message. The first post starts the thread; this may be called the *entry*. Posts that follow in the thread are meant to continue the discussion about the *entry*, or respond to other replies. A post mainly contains four parts: the author, the title, the time-stamp and the content.

Thread: A thread is a collection of posts. Each post belongs to and only belongs to one thread. A thread can contain any number of posts, including multiple posts from the same member, even if they are one after the other.

Topic: A topic on BBS is formed with one or more threads.

* Supported by the National Science Foundation of China under Grant No.60873134.

[1] http://www.readwriteweb.com/

W. Liu et al. (Eds.): WISM 2009, LNCS 5854, pp. 129–138, 2009.

Nowadays, to get hot topics on BBS, the most popular method is to see Top-N topics based on the number of page views and replies each day, such as “Today’s TOP HOT” in SMTH[2] BBS . However, this method may not be a suitable solution to user’s concern due to its limits. The users usually want to have an outline of discussions to see “what’s hot now” or “what’s hot today”.

Hot Topic Detection algorithm proposed in this paper is a good practice of topic detection and tracking [1](TDT) on BBS. We find some common features on BBS: Firstly, language used in BBS is more informal, colloquial and with many abbreviations. Even those experienced users could not know all these abbreviations. Secondly, it is very common using emotional signs like “!!!”, “^_”, “re”. Some posts include nothing but those signs. Thirdly, a single thread may reflect different topics. A survey thread like “how are you going on these days”, may trigger other topics during the discussion. In addition, unlike traditional news reports, the information of users’ participation is helpful to find hot topics on BBS.

The remainder of this paper is organized as follows: In Section 2, we discuss some related work on hot topic detection. In Section 3, we first define the hot topic, and introduce the aging theory. Then we describe our approach for hot topic detection on BBS using aging theory. In Section 4, we discuss the results of the experiments run. Finally, in section 5, we present our conclusions and some future research directions.

2 Related Work

Topic Detection and Tracking (TDT) has been researched for a very long time [1] [2] [3], however, former researches on hot topic detection are mainly based on those traditional, formal news reports [5] [6] [7]. If these methods are applied to BBS directly, it will surely lead to bad results [4]. With the rapid development of social networking recent years, researchers get to focus on topic detection towards social media such as BBS, BLOG, etc [4].

In this paper, we try to find hot topics on BBS, which shows a high threaded discussion level with strong interactivity and many informal terms. In 2005, Lan You elt [8] did a similar research to find hot topics on BBS. In his method, threads were first clustered into topics based on their lexical similarity. Then a BPNN(Back-Propagation Neural Network) based classification algorithm was used to judge the hotness of topics according to their popularity, quality as well as thread distribution over time. Although this method can find hot topics in a certain period of time, it would be better if we enhance the influence of hot topics at the observation point.

In addition, BBS is a social network, its interaction often impacts the generation of hot topic. Hot topic, is always accompanied by the abundance of users’ participation. Researchers who study the social networking have already found some interesting results [9] [10]. There is a simple but effective way to analyze the relationship of posts and measure the importance of them: *out-degree* and

[2] http://www.newsmth.net/

in-degree. It's very helpful to improve the topic model [11] using this way. *out-degree* means how many posts this post replies to and *in-degree* means how many posts reply to this post.

3 Hot Topic Detection on BBS Using Aging Theory

Although daily posts involve a lot of topics, more than 60% of the posts just focus on a few topics. Therefore, it is very meaningful to find hot topics on BBS.

3.1 Definition of Hot Topic

To find out hot topics, we need to define the hot topics. There are four distinct characteristics of "hot topic":

Massive Posts: Only an attractive topic can assemble lots of users' discussion, which, in turn, becomes a prerequisite for a hot topic. This factor is comprised in our energy calculation process. Each post contains certain nutrition which can be transformed to the energy, therefore, topics with more posts could gain more energy.

High Quality Posts: Compared with those junk posts (like "I agree", "good"), a hot topic always has more posts of high quality. The relationship among posts could help us to identify which posts have high quality.

High Cohesion: Since scattered content has less attraction, the content of a hot topic is usually compact and centralized. We use the threshold of Single-Pass clustering to strictly control the number of threads to form the topic.

Bursting: For a hot topic, it often gathers a large number of posts in a short period of time, and then gets to a stable state until slowly disappear, which implies a life cycle of the topic.

3.2 Aging Theory

We regard the Rise and Fall of the topic as a life cycle. Chen [12] is the first person who models the topic using aging theory in 2003. He divides the life cycle into four stages: birth, growth, decay and death. To track life cycles of topics, aging theory uses the concept of energy function. The value of energy function shows how active a topic is. The energy of a topic increases when the topic becomes popular, and diminishes with the time. If we use the concept of nutrition, each post can be seen as a food to the topic, which contain certain nutrients, and these nutrients can be transformed to the energy value using energy function.

Time line can be equally divided into time slots(we take 3 hours as a time slot in our experiment),and then we employ three functions of aging theory to update the energy value of the topic at the end of each time slot.

getNutrition(): Calculate how much nutrition a topic can get from posts.

energyFunction(): This is a monotonically increasing function. It can transform the nutritional value into energy value. $energyFunction^{-1}()$ transforms the energy values into the nutritional value of the topic.

energyDecay(): In the end of each time slot, we perform the attenuation of energy to the topic. If the energy value declines below β, this topic will be moved to removed list, and not used for following steps. β is a decay factor, which can be obtained from the training data.

3.3 Hot Topic Detection on BBS

After given a observation point, we start to analyze the discussion data several time slots before the point. Our method will perform in three steps: (1) Candidate Topic Discovery. We use incremental Single-Pass clustering method to get the candidate topic list in each time slot. (2) Topic energy calculation. In this step, it will give each topic an energy value at the end of each time slot. Both two steps above are carried out incrementally.(3) Hot topic ranking. We rank the topics according to their energy value and then verify them whether or not to meet our definition of hot topic. We will finally get the list of hot topics for the observation point. What's more, according to its energy value curve, we can get a clear picture of each topic's rise and down.

Candidate Topic Discovery. We try to get the candidate topics in the time slot. We set the currently topic set as T, which initialized as $NULL$. Then we use the following method: determine which topic in topic set T is the most similar with the new thread d in current time slot. If their similarity exceeds the predefined $threshold_{sim}$, we merge the thread d into this topic and update the topic vector, otherwise this thread d is considered as a new topic t', and put t' into the topic set T. When all of the threads in current time slot have been dealt, we will get the topic set of the current time slot T. $threshold_{sim}$ needs to be identified in the training data.

Topic Energy Calculation. After candidate topics are discovered in the time slot, each topic should be given an energy valule.

First, we need to measure how much nutrition each post contains.High quality posts usually have more nutrition, which means it can offer more energy. The nutrition of a post is calculated by the Formula 1:

$$getNutrition(p) = z \cdot \frac{\lg(y+2)}{\lg(y+2)} \tag{1}$$

where
$getNutrition(p)$: indicates the nutrition the post p contains
z: denotes the content similarity between the post and the topic
y: denotes the reference number by others posts

Secondly, for all new generated posts in the time slot of the topic, calculate their nutrition and transform the nutritional value into the energy value, then add it to the cumulative energy value of the topic.

At the end of each time slot, we will get a cumulative energy value. It can be divided into two parts. One is the energy value of the last time slot. The other is energy variation in the current time slot. There are a 3-step calculation process:

1. Use $energyFunction^{-1}(e_t^{i-1})$ to convert the topic t 's energy value e_t^{i-1} of last time slot i into nutrition value n_{t-1}.
2. Get the nutrition from all the posts of the topic in the time slot, which noted as n_d (Formula 2), and add it to the legacy value n_{t-1} using α as nutrition transferred factor which is defined in the aging theory (Formula 3). α need be identified from the training data.

$$n_d = \sum_{p \in d}^{p.time \subseteq i} getNutrition(p) \tag{2}$$

$$n_t = n_{t-1} + \alpha \cdot n_d \tag{3}$$

3. Calculate the cumulative energy value at the end of current time slot, using the formula: $energyFunction(n_t)$. The specific form of $energyFunction()$ is defined as Formula 4, analogous to that used in [12]:

$$energyFunction(x) = \begin{cases} \dfrac{1}{1+x}, & x \geq 0, \qquad (4) \\ 0, & x < 0. \qquad (4') \end{cases}$$

Thirdly, perform energy decay on topics. Energy of topics increases with upcoming posts adding to topics, but also decays with the time passing by. The topic energy gets modified by a decay factor that represents the decay in each time slot. If in a certain time slot, the energy value which the topic obtains is less than that it decays, then the topic will show signs of recession. When the energy of a topic is below the predefined threshold, it is supposed to be in a "death" state. So we remove it from the set of survival topics T and add it to the set of removed topics T_{rem}. It will not used in the following steps.

Hot Topic Ranking. We get hot topic candidates by rank the topics according to their energy value. Then, according with the definition of the hot topic, it should have the feature of bursting. In our experiment, if a topic lasts for more than 24 hours, and its standard variation of the energy value is less than 0.05, this topic is considered as no bursting. These topics are most probably composed with very common threads titled of "how about. . . , how to get to . . . ", which are frequently appeared all day long. After filtering out the topic whose standard variation is less than 0.05, we get hot topics with highest energy value.

4 Experiment Analysis and Result

In order to evaluate the proposed method, we make experiments on the BBS data.

4.1 Corpus and Parameter Settings

At present, there is no public BBS corpus, so we create a crawler to get the daily posts in the board "NewExpress" on SMTH BBS which is one of most popular BBS in China. System notifications on BBS are ignored by our crawler. All the data were published from March 14,2009 to March 30,2009. There are up to 14,807 threads which totally including 74,506 posts. In average, there are 871 threads including 4,382.7 posts per day. All posts have been already preprocessed.

The corpus was divided into two data sets:

Training Set: Published from March 14 to March 21, which contains 6,903 threads, 34,036 posts, and 10 topics are manually labeled.

Test Set: Published from March 22 to March 30, which contains 7,904 threads, 40,470 posts.

4.2 Hot Topic Detection Experiment on BBS

Hot topics on BBS usually last for a short-term period and are more concentrated. BBS users concern more about hot topics at the point when logging on to BBS. Therefore, it is worth-while to find out accurate hot topics happened recently. To the best of our knowledge, there is no general standard for hot topic evaluation on BBS. Here we adopt "Top-5 Hot Topics" issued by original BBS board as a "*BaseLine*" to measure our results. The experiment is divided into two parts, respectively, hot topics detection on one day and on three days.

Before the experiment, all parameters are settled optimized according to the training data. After validation, it gets to the best result when $threshold_{sim}$ equals 0.2213. α, β are determined according to the method used in [13]: For each topic, a proportion r_1 of the total nutrition corresponds to a proportion s_1 of the total energy. α, β can be determined using two points(r_1,s_1), (r_2, s_2). At last, the final parameter values are the averages of α, β of 10 topics: α=0.118659, β=0.079687. Then the threshold for the standard variation of energy value V is set as 0.05. In addition, each time slot is represented as 3 hours and the observation point is the last minute of each day.

Hot Topic for One Day: Firstly, we perform our experiment to detect hot topics each day during March 22 to March 30. The experimental result is validated by "*Baseline*" of each day. For convenience, the result generated using our method will be noted as "*RUA*"(Result Using Aging theory).

Other symbols used in the result column are predefined as follows:

➠:Hot topic generated by *RUA* is almost equal to the *Baseline* at the same rank.

$\Uparrow$: Hot topic discovered by both *Baseline* and *RUA*, but this topic ranks higher in *RUA*.

$\Downarrow$: Hot topic discovered both by *Baseline* and *RUA*, but this topic ranks lower in *RUA*.

Table 1. Daily top-5 hot topics on BBS detected on March 26. Comparative results between the *RUA* and *Baseline* are shown at Result column "Re.".Topics are listed in the descending order of hotness.

March 26			
Topic	*RUA*	*Baseline*	Re.
1	①Be a Buddhist monk is not a job(13) ②Where are Buddhist nun from?(30) ③How hard to be vegetarians as Monks(15) ④Why are some Monks married?(31)	①What can you guess is the name of first Chinese Aircraft carrier?	**N**
2	①What can you guess is the name of first Chinese Aircraft carrier? ② Name the first Aircraft carrier ③Original concept of Aircraft carrier ④ China would build Aircraft carrier? ⑤The first Chinese Aircraft carrier should be named as 'Never supremacy'	① ShingTung Yau, a big lie?□	**M**
3	①Haizi's in a low status in poetry community ② 20th anniversary for Haizi ③20 years after Haizi Suicide	① Some middle school in Beijing,very bossy□	**M**
4	①Tianqiao Chen and Yuzhu Shi, who is the best	①Turned down for many times,But see my love today	⇑
5	①Turned down for many times, But see my love today	① ShingTung Yau, no logical?□	⇓

N(ew): Hot topic generated by *RUA*, but no corresponding topic appears in *Baseline*.
M(erge): Hot topic in *RUA* merges two or more topics in *Baseline*.
□ : Hot topic appears in *Baseline*, but not in *RUA*.

The result on March 26 is listed in Table 1.Thread titles are listed to represent certain topics. Energy variation for each topic is also shown in Figure 1. The rest result of other 8 day is summarized in the following analysis.

We find three interesting discoveries through this experiment. First of all, *RUA* has the basic coverage of hot topics which retrieved by *Baseline*. These marked with ➠, ⇑, ⇓, or **M** mean topics are both detected by *Baseline* and *RUA*, only in different ranks or different styles of organization. After manual analysis, among 45 topics retrieved in 9 days, there are 35 topics are labeled with marks above, accounting for 77.8% of all. Moreover, there are 5 in above 35 topics marked with '**M**', that means *RUA* can not only find the topic, but also be able to enhance the coherence of the topic by associated with related threads of the topic. This will show the outline of the topic for users. For instance, in result on March 26 showed in Table 1, the second topic detected by *RUA* not only found the hot topic "What, can you guess, is the name of first Chinese Aircraft carrier?", but also organized threads which related to this topic such as "The first Chinese Aircraft carrier should be named as Never supremacy!"etc.

Secondly, *RUA* is competent for detecting hot topics which can not obtained by *Baseline*. It's known that the discussion of a topic may be scattered in a

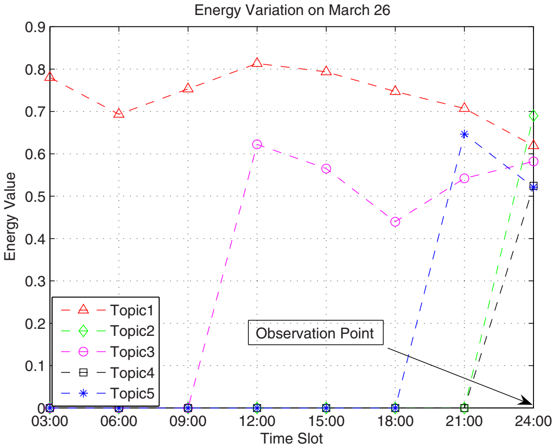

Fig. 1. Energy distribution of daily hot topics on BBS detected by RUA on 26 March

number of threads, if none of these threads gather abundant posts, the $Baseline$ method can hardly find it. Instead, RUA can find these topics(which labeled '**N**' in Table 1) such as topic 1 in Table 1. Though the posts number of each thread is not too much (the posts number is shown after each thread title), accumulation of all these posts reaches up to a large number. From the corresponding energy variation in Figure 1, we know that this topic just starts to thrive in a short period of time. According to the definition of hotness, it is believed as a hot topic. There are 5 in 45 topics marked as '**N**', accounting for 22.2%. Except one topic, 4 other topics are believed as hot topics through manually validating.

Finally, RUA is capable of finding more time sensitive topics. Some topics labeled as '□', are detected in the $BaseLine$, but do not appear in RUA, such as topic 2 of $BaseLine$ in Table 1. After looking into all the threads of this topic, we know that it was discussed a lot before 12:00 on the 26th, but received little concern afterwards. Therefor, it is not suitable to be a hot topic for our observation point at the last minute on the day. RUA ranks the energy at the observing point which ensures the timeliness of hot topics.

Hot Topic for Three Days. We also perform an experiment to detect hot topics for three days long during March 22 to March 30. The experimental result on the time period of March 28 to March 30 is shown in Table 2, and corresponding energy variation in Figure 2. The other result are not listed here due to length of the article. Through this experiment, we find that RUA can also effectively detect hot topics which last for a relatively long time period. For example, in Figure 2, topic 1 which discussed about holiday policy of Labor Day happened one and half a day ago, and followed with a period receiving little attention. Then it got to be ultimately culminated. Those topics last such long time are hardly detected by $BaseLine$ method.

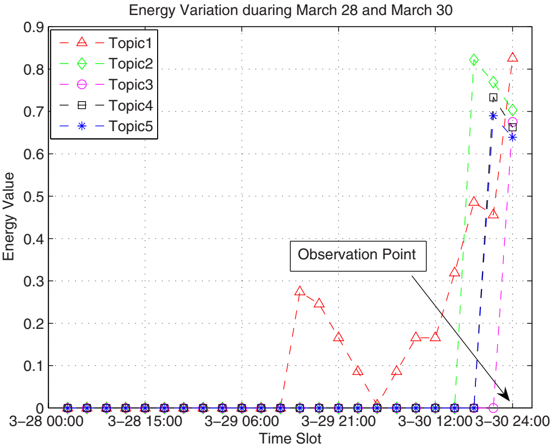

Fig. 2. Energy distribution of hot topics on BBS during 22 March to 30 March

Table 2. Hot topics on BBS of three days long during 28 March to 30 March.Topics listed in descending order of energy value.

March 28 to March 30	
Topic	Experimental Result
1	① NEWS:Guangdong,Cancel the Labor Day Plan!! ②Recover the Golden Week of Labor Day in Guangdong ③Guangdong Government is pig Brain... ④ Professor Cai's Holiday policy report
2	①Is it good to be a teacher at Tsinghua? ②Who can tell me how to get along with students? ③Professor Cai jumped from NaiKai univ.
3	①If civilian workers give up farming? ②Finally understand the reason of farmer's low payment
4	①Let's see CCTV8, LiZhu jump on Xiaoru Fang!
5	Who is a handsome scholar or a pretty girl worldwide?

5 Conclusion and Future Work

Based on aging theory, our method can find the hot topics in any period of time, such as hot topics of the day, hot topics of 3-days and so on. It can improve hot topics retrieval on BBS. Users can know what's going on clearly and quickly.

The main contributions of the paper are as follows. First, we analyze the characteristics of hot topic in BBS. Second, we propose a effective way to validate the post importance. We apply aging theory on BBS posts in order to get the hot topics. After comparison with the method in [8] which employs quite a few parameters to measure the hotness of the topic, we summarize our method: (1) With the help of energy calculation, many spam posts are avoided. (2) Large training data is not needed. It saves human labor. (3) The choice of the observation time point will have an impact on hot topics detected by our method.

Our next step will focus on how to combine BBS, Blog, even Twitter[3] to find hot topics.

References

1. Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y.: Topic detection and tracking pilot study: Final report. In: Proc. of the DARPA Broadcast News Transcription and Understanding Workshop. OMG Press, Needham (1998)
2. Canhui, W., Min, Z., Liyun, R., Shaoping, M.: Automatic online news topic ranking using media focus and user attention based on aging theory. In: Proceeding of the 17th ACM conference on Information and knowledge management, pp. 1033–1042. ACM Press, New York (2008)
3. Yang, Y., Pierce, T., Carbonell, J.: A study of retrospective and online event detection. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 28–36. ACM Press, New York (1998)
4. Zhu, M., Hu, W., Ou, W.: Topic Detection and Tracking for Threaded Discussion Communities. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 77–83. IEEE Press, Washington (2008)
5. Huimin, Y., Wei, C., Guanzhong: Design and implementation of online hot topic discovery model. J. Wuhan University Journal of Natural Sciences 11(1), 21–26 (2006)
6. He, T., Qu, G., Li, S., Tu, X., Zhang, Y., Ren, H.: Semiautomatic Hot Event Detection. In: Xue, J., Osmar, L., Zhanhuai, R., Xi'an, L, eds. (2006)
7. Chen, K.Y., Luesukprasert, L., Chou, S.c.T.: Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling. In: IEEE Transactions on Knowledge and Data Engineering, pp. 1016–1025. IEEE Press, Piscataway (2007)
8. Lan, Y., Yongping, D., Jiayin, G., Xuanjing, H., Lide, W.: BBS based Hot topic retrieval using backpropagation Neural Network. In: Su, K.-Y., Tsujii, J., Lee, J.-H., Kwong, O.Y. (eds.) IJCNLP 2004. LNCS (LNAI), vol. 3248, pp. 139–148. Springer, Heidelberg (2005)
9. Robert, D.N., Lina, Z.: Social Computing and Weighting to Identify Member Roles in Online Communitie. In: Proc. of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (2005)
10. ZhiLi, W., Chunhung, L.: Topic Detection in Online Discussion Using Nonnegative Matrix Factorization. In: Proc. of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology Workshops (2007)
11. Tuulos, V., Tirri, H.: Combining Topic Models and Social Networks for Chat Data Mining. In: Proc. of the 2004 Web intelligence International Conference, pp. 206–213. IEEE Press, Washington (2004)
12. Chen, C.C., Chen, Y.T., Sun, Y., Chen, M.C.: Life Cycle Modeling of News Events Using Aging Theory. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) ECML 2003. LNCS (LNAI), vol. 2837, pp. 47–59. Springer, Heidelberg (2003)

[3] http://www.twitter.com/