Online High-Quality Topic Detection for Bulletin Board Systems

Jungang Xu†1, Member, Hui Li†, Yan Zhao††, and Ben He†, Nonmembers

SUMMARY Even with the recent development of new types of social networking services such as microblogs, Bulletin Board Systems (BBS) remain popular for local communities and vertical discussions. These BBS sites have high volume of traffic everyday with user discussions on a variety of topics. Therefore it is difficult for BBS visitors to find the posts that they are interested in from the large amount of discussion threads. We attempt to explore several main characteristics of BBS, including organizational flexibility of BBS texts, high data volume and aging characteristic of BBS topics. Based on these characteristics, we propose a novel method of Online Topic Detection (OTD) on BBS, which mainly includes a representative post selection procedure based on Markov chain model and an efficient topic clustering algorithm with candidate topic set generation based on Aging Theory. Experimental results show that our method improves the performance of OTD in BBS environment in both detection accuracy and time efficiency. In addition, analysis on the aging characteristic of discussion topics shows that the generation and aging of topics on BBS is very fast, so it is wise to introduce candidate topic set generation strategy based on Aging Theory into the topic clustering algorithm.

key words: online topic detection, BBS, representative post selection, candidate topic set, clustering

1. Introduction

The rise of social networking services has created a highly dynamic Web communication medium that attracts many grassroots Web users. While microblogging Web sites such as Facebook and Twitter are becoming popular with more than 100 million active users every month [1], [2], Bulletin Board System (BBS) remains an important means of online communication, especially for vertical discussions on specific topics, e.g., Linux Operating Systems, Digital Cameras or local events. For example, Hiapk, a Chinese vertical forum for Android discussions, has more than 10 million members currently [3].

The information volume of BBS is pretty large and the topics in BBS discussion show scattered features. Although most BBS sites are organized by categories, many posts which should belong to the same topic often scatter in different categories, so it is difficult for visitors to find topics among multiple categories. According to the report from CNZZ in 2010 [4], even the most popular topic in BBS has only attracted 3.1 percent of visitors. It has become more and more difficult for BBS visitors to retrieve interesting topics, so it is a meaningful and challenging work to research how to detect topics from the great amount of discussion texts of BBS.

Online Topic Detection (OTD) for BBS aims to detect new topics and find topic-related texts from continuously incoming BBS posts for visitors. As a subtask of Topic Detection and Tracking (TDT), there have been many research efforts on OTD in the last decade [5], [6]. Previous approaches for OTD/TDT are mostly based on the clustering algorithms and dedicated to news web pages. Compared to previous approaches, OTD for BBS faces new challenges on both detection accuracy and time efficiency. Because of high demand of time efficiency, it must be possible to execute the detection algorithm incrementally.

Before the introduction of OTD for BBS, there are some definitions of the basic elements of BBS structure.

Post: each piece of text submitted to BBS by a user. If a post is the first one of the current discussion, it is called an entry post. If postp is a reply to postx, then postp is the child post (or reply post) of postx, and postx is the parent post of postp.

Thread: a container for all the posts from the same discussion. We assume that all the posts in a thread share the same headline.

Compared to traditional TDT corpus, such as news report texts published by Linguistic Data Consortium [7], the characteristics of BBS texts can be summarized as follows.

Firstly, BBS texts are organized flexibly in threads. Each thread is composed of multiple posts, but each news report consists of only one single text in most cases. The structure of BBS texts is similar to people’s discussion. With the discussion going further, the subsequent posts of the thread are filled with many divergent trivial things and opinion expressions which may have little relation with the original topic.

Secondly, the quantity of BBS texts is much larger than that of news web pages, because BBS texts are contributed by all BBS users while news web pages are only maintained by the site editors. So there is a high demand for efficiency of online topic detection algorithms for BBS.

Lastly, generation and aging of the topics on BBS are very fast. This characteristic is a very important inspiration to our topic clustering algorithm as discussed in Sect. 5.

Based on the characteristics of BBS texts listed above, we propose a complete method of OTD for BBS, which mainly improves the existing OTD method in the following two aspects: (1) proposing a representative post selection
strategy based on Markov chain model to extract representative texts of threads; (2) introducing a candidate topic set generation strategy into clustering procedure of OTD based on Aging Theory to improve the efficiency of the topic clustering.

Our method is evaluated on both accuracy and efficiency. The experiment platform is the BBS called Newsmth[8] which is a famous BBS forum in China with about 30,000 online members, more than 10 discussion boards and 25,000 posts every day.

The rest of the paper is organized as follows. In Sect. 2, we briefly review the related work. In Sect. 3, we introduce our OTD algorithm on BBS. In Sect. 4, the experimental results on a real BBS forum dataset are presented. In Sect. 5, we offer a discussion on generation and aging features of BBS topics. Finally, in Sect. 6, we summarize our findings and suggest possible directions for future work.

2. Related Work

There have been quite a few studies on OTD in the last decade. While most of the existing methods are on OTD for news web pages, OTD methods for BBS are also addressed in the studies mentioned below.

In the existing researches on OTD, clustering algorithms based on text similarity have achieved great success. Y. Yang et al[5] proposed two algorithms, a hierarchical algorithm based on group-averaged clustering and a Single Pass Clustering algorithm, which can automatically detect novel events from a temporally-ordered stream of new stories. J. Allan et al[6] used Single Pass Clustering algorithm[9] and a novel thresholding model to detect and track events within a stream of broadcast news stories. Among all these clustering algorithms, Single Pass Clustering algorithm is considered to achieve the best time performance[10]. Compared to news web pages, BBS texts are more complicated and filled with user participation. M. Zhu et al[11] proved that some methods doing well with news web pages do not achieve good performance in BBS environment. According to the differences between BBS and news web pages, a few research results in BBS environment are published. J. Kim et al[12] proposed a method to split the topics in a single thread by organizing the posts in a thread as a chain and calculating the similarity between consecutive posts in the chain. However, posts from different threads cannot be organized as a chain, so this method is not appropriate for detecting topics from a set of threads. D. Zheng and F. Li[13] introduced Aging Theory into BBS topic ranking, but their approach may not be optimal as it does not adequately utilize the structure of threads to select representative posts. However, this aspect has a significant impact on the detection accuracy. N. Wanas et al[14] proposed an approach for off-topic detection in threads based on keyword extraction, which considers more about the probability distribution of the words, but ignores the reply relations and visiting paths of the visitors. M. Zhu et al[11] extended the basic TDT framework with a user activity model UF-ITUF. These methods achieved great success, but did not take all characteristics of BBS into consideration overall.

Our work aims to propose an OTD method for BBS, which includes a representative post selection strategy while preprocessing and an efficient topic clustering algorithm. Different from those clustering algorithms above, our topic clustering algorithm is specially optimized for BBS by utilizing topic vitality information. In our work, we also use vector space model and incremental TF-IDF model which are widely used in TDT[15]–[18].

In addition, the representative post selection strategy in our work is inspired by Pagerank algorithm proposed in[19] for Web search. Our work is different from that in Pagerank because the reply relationship between posts is in fact a logical connection, which does not only depend on hyperlinks.

3. OTD Method for BBS

3.1 Overview

Here, based on the characteristics of BBS listed in Sect. 1, we propose an OTD method specifically for BBS. First, a representative post selection strategy based on the Markov chain model is adopted to filter out the noisy and irrelevant posts, which makes great contribution to the accuracy of our topic clustering algorithm. In addition, candidate topic set generation strategy based on Aging Theory is introduced into our algorithm to improve the efficiency of the topic clustering algorithm.

The main process of our proposed OTD method for BBS includes representative post selection, text feature vector generation and topic clustering. The first step, representative post selection, which removes low-quality and off-topic posts on a per thread basis. The second step, text feature vector generation, which transforms representative texts into thread vector by incremental TF-IDF model. The last step, topic clustering, which clusters thread vectors with similar topic using our topic clustering algorithm. To improve time efficiency of the algorithm, we use Aging Theory to minimize the size of candidate cluster set.

3.2 Representative Post Selection

A BBS thread with multiple posts can be represented by a tree-like structure as shown in Fig. 1.

The whole tree represents a thread. Each node in the tree represents a single post in the thread. Each link between two tree nodes represents the reply relations between two posts.

With the discussion in a thread going further, more and more posts are posted. The subsequent posts of the thread are usually filled with many trivial posts and personal opinion expressions which may have little relation with the topic. These low-quality posts can bring noise into topic detection, so a representative post selection procedure is needed in OTD method for BBS to pick out the high-quality posts.
A post is of high quality if it can attract enough user attentions even though it is not like the entry post. That’s because the topic of a thread depends on the posts that users are interested in but not the entry post. Because the reply relation of the posts is a direct reflection of user attention in BBS, each reply post can be considered as an approver of post quality for its parent post.

Usually, a high-quality post has more reply posts than a low-quality post. But it is not sufficient to judge the quality of a post only by the number of reply posts, such as the example shown in Fig. 2. In Fig. 2, Post C and Post D both have 5 derivative posts, but the derivative posts of Post D are more related with Post 4 than Post D itself, so the quality of Post C should be higher. On the other hand, the number of reply posts should include not only direct reply posts but also indirect reply posts, such as the example in Fig. 3. In Fig. 3, although Post A and Post B both have 3 direct reply posts, Post A is more important than Post B because it has more indirect reply posts obviously.

The critical problem of the representative post selection is how to measure the quality of a post. Through two examples above, we can also observe that it is possible to rank the quality of the posts in a thread by reply topology, which is a novel way to identify low-quality posts in thread. The ranking problem of post quality is quite similar with page ranking problem in search engine. Inspired by Pagerank algorithm [19], we propose the following model for selecting high-quality posts from a given thread.

Suppose that there is a BBS visitor who wants to figure out the most important topic of the discussion within a thread s. This visitor would browse those posts through reply relation among posts. After she or he reads a certain post, the next post she or he is going to read is either one of the child posts of the current post or the parent post. This behavior is reasonable because people tend to read coherently, and a reply implies a relation between these two posts.

Suppose that the reading transfer behavior of BBS fits Markov model, the next reading post only depends on the current post and has no relation with the past reading behaviors. In this model, each behavior of reading certain post is a Markov state. The transition probability between different Markov states can be measured by the following hypothesis.

Let α be the probability of taking the parent post as the next reading material (α is called reading transformation factor in our model, which is trained from manually ranked posts). In our model, every child post of the current post is treated equally, so if the current post has $C_i$ child posts, then the probability of reading the next certain child post is $(1 - \alpha)/C_i$. Specially, when the current post has no parent posts or child posts, such as the entry post of the thread, the corresponding probability should be shared by all posts in the current thread.

Then the probability of taking post $j$ as the subsequent reading material of post $i$ can be obtained by solving the Eq. (1), Eq. (2) and Eq. (3).

\begin{align*}
(1) & \quad P_{ij} = \begin{cases} 
\frac{1 - \alpha}{C_i} + \frac{\alpha}{N}, & j \text{ is } i's \text{ child } \\
\frac{\alpha}{N}, & \text{otherwise}
\end{cases} \\
(2) & \quad P_{ij} = \begin{cases} 
\frac{\alpha + 1 - \alpha}{N}, & j \text{ is } i's \text{ parent } \\
1 - \frac{\alpha}{N}, & \text{otherwise}
\end{cases} \\
(3) & \quad P_{ij} = \begin{cases} 
\alpha, & j \text{ is } i's \text{ parent } \\
\frac{1 - \alpha}{C_i}, & j \text{ is } i's \text{ child } \\
0, & \text{otherwise}
\end{cases}
\end{align*}

Where $P_{ij}$ denotes the probability of taking post $j$ as the subsequent reading material of post $i$. $N$ represents the number of posts in the thread. $C_i$ represents the number of child posts of post $i$.

This Markov chain is a typical ergodic Markov chain. According to the principle of Markov chain, there is a
unique steady-state probability vector $\bar{\pi}$, which is the left eigenvector of $P$ as shown in Eq. (4).

$$\bar{\pi} = \pi \cdot P$$

Where $pv_k$ denotes the steady probability of Markov state $k$. In our model, $pv_k$ also represents the ratio of time that the users spent on post$_k$. Therefore, the larger $pv_k$ is, the more attention post$_k$ is, has been paid, which usually means that the quality of post$_k$ is better and the probability of post$_k$ representing the core topic of the current thread is higher. So the quality of post $k$ can be represented by $pv_k$, too.

Therefore, we solve the ranking problem of post quality through solving Eq. (4). Our strategy for representative post selection is to choose $m$ posts with the highest $pv$ value, where $m$ is an empirical parameter. In our experiment, $m$ is set to 7.

For example, there is a thread as in Fig. 4. Let $\alpha = 2/3$, which means that the probability of transferring to read parent post of the current post is 2/3. Then the transition probability matrix of reading is as follows.

$$P = \begin{bmatrix}
\frac{2}{13} & \frac{3}{10} & \frac{3}{10} & \frac{2}{13} & \frac{2}{13} \\
\frac{3}{10} & 0 & 0 & \frac{2}{13} & \frac{2}{13} \\
\frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} \\
\frac{1}{13} & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} \\
\frac{1}{15} & \frac{1}{15} & \frac{1}{15} & \frac{1}{15} & \frac{1}{15}
\end{bmatrix}$$

By solving Eq. (4), we get the static probabilities, $\bar{\pi} = \{0.232, 0.168, 0.110, 0.245, 0.245\}$. The quality sequence of all the posts is: post$_3 =$ post$_4 =$ post$_0 =$ post$_1 >$ post$_2$. So if we choose the 3 highest posts as the representative of the current thread, the selected posts should be {post$_3$, post$_4$, post$_0$}.

After the representative post selection, a few high-quality posts are combined as a pseudo document. According to bag-of-words strategy, each thread is represented as a $|\nu|$-dimensional vector where each dimension is TF-IDF feature of a unique term in the dictionary. The method of text feature vector generation is described in the following subsection.

### 3.3 Text Feature Vector Generation

After representative post selection, each BBS thread is represented as a pseudo document of bag-of-words composed of high-quality posts. In this procedure, text feature vectors of the threads are generated by incremental TF-IDF model, which is widely used in traditional TDT [15]–[18].

The global Document Frequency (DF) vector is initialized by combined texts of the threads in the training set. For each new thread, DF vector is refreshed as in Eq. (5).

$$df_d(w) = df_{d-1}(w) + df_{d\Delta}(w)$$

Where $df_d(w)$ denotes the document frequency for word$_w$ after the processing of thread$_d$; $df_{d\Delta}(w)$ is equal to 1 if word$_w$ exists in thread$_d$, 0 otherwise.

Because words in the headline are considered to be more valuable than those in the content, we slightly adjust the weight of the words in different positions as in Eq. (6).

$$tf(d, w) = r \times tf_{\text{headline}}(d, w) + tf_{\text{content}}(d, w)$$

Where $tf(d, w)$ is the term frequency of word$_w$ in thread$_d$; $tf_{\text{headline}}(d, w)$ is the term frequency in the headline while $tf_{\text{content}}(d, w)$ is the term frequency in the content; $r$ is the weight of words in the headline which is a tunable parameter trained in the first experiment called performance evaluation in accuracy in Sect. 4.3.

The final weight of a word is calculated by combining $tf$ and $df$ vector as in Eq. (7).

$$weight(d, w) = \frac{1}{\text{norm}(d)} tf(d, w) \cdot \log \frac{N_d}{df_d(w)}$$

Where $N_d$ denotes the count of threads up to thread$_d$; norm$(d)$ denotes the normalization factor. Because cosine distance is used for similarity computation in clustering, the normalization factor norm$(d)$ is calculated as in Eq. (8).

$$\text{norm}(d) = \sqrt{\sum_w \text{tf}(d, w) \cdot \log \frac{N_d}{df_d(w)}}$$

After text feature vector generation, each thread is projected into a $|\nu|$-dimension vector. Every dimension of the vector represents the weight of the corresponding word.

Similarity between two vectors is calculated by cosine distance as in Eq. (9).

$$\text{dis tan ce}(d_1, d_2) = \frac{\sum \text{weight}(d_1, w) \times \text{weight}(d_2, w))}{\text{norm}(d_1) \times \text{norm}(d_2)}$$

Cosine distance is used to measure similarity between two text feature vectors. The larger cosine distance is, the more similar two vectors are.

### 3.4 Topic Clustering with Candidate Topic Set

The core idea of topic detection algorithms based on text similarity is that the more in alignment of topics and the less in distance between text feature vectors. Online topic detection problem can be converted into a pure clustering problem.

Single Pass Clustering algorithm is a widely used topic
there are \( \mu \) the time cost of the first \( n \) threads is as in Eq. (11).

\[
\sum_{m=1}^{n} (m-1)\mu = \frac{1}{2}\mu(n^2 - n) \tag{11}
\]

So the actual time complexity of Single Pass Clustering in OTD is \( O(n^2) \). Although Single Pass Clustering has almost the lowest time complexity among all the topic clustering algorithms [10], it is still too slow when it comes to large datasets.

In order to measure the vitality of the topics, we introduce Aging Theory [15] into Single Pass Clustering.

Aging Theory can be used to model the life circle of one topic. The life circle of a topic is similar to the aging procedure of creatures: birth, growth, decay and death. In Aging Theory, the vitality of a topic is represented by the energy. When topic-related posts show up, the topic energy rises. The energy value is continuously decreasing over time. When the energy value of a topic reaches 0, the topic becomes dead and can never be used again.

Each topic-related post contains certain nutrition which can be transformed to topic’s energy. The nutrition value is calculated as in Eq. (12).

\[
\Delta N_t(s) = \beta \tag{12}
\]

Where \( \Delta N_t(s) \) denotes the nutrition value that post \( s \) contains, \( \beta \) denotes the nutrition of each post. In our experiment, \( \beta \) is simply set to 1.

The relationship between energy and nutrition can be represented by a sigmoid function. In our research, the energy of a topic can be calculated from nutrition value \( x \) using the following \( EF \) (abbreviation of Energy Function) as in Eq. (13).

\[
EF(x) = \begin{cases} 
\frac{\lambda x}{1 + \lambda x}, & x \geq 0 \\
0, & x < 0 
\end{cases} \tag{13}
\]

Where \( \lambda \) is an empirical parameter, which is always less than 1. In our experiments, \( \lambda \) is set to 0.25.

The refreshing of topic energy is carried out periodically. Because the incoming speed of the new posts is not constant (such as, faster in the evening than that in the morning), the period of refreshing is not measured by time, but by the number of incoming posts.

In our research, time line is divided into some slots, each slot contains the same number of new posts. In each slot, the following energy refreshing procedure is carried out. The whole procedure consists of two steps, energy boost and energy decay.

In the step of energy boost, the energy of each active topic cluster is boosted by topic-related posts. The calculation of energy boost in slot \( t \) is as in Eq. (14).

\[
\text{eng}'(k) = EF\left(\frac{1}{EF^{-1}(\text{eng}_{t-1}(k))} + \sum_{s\in TS_k} \Delta N_t(s)\right) \tag{14}
\]

Where \( \text{eng}(k) \) is the energy of cluster \( k \) in slot \( t \), \( TS_k \) represents the union of all posts belonging to topic cluster \( k \) including the representative posts and non-representative posts.

After the step of energy boost, the step of energy decay is carried out. The energy of each active topic cluster keeps decaying with a constant ratio over slots. The calculation of energy decay is as in Eq. (15).

\[
\text{eng}(k) = \begin{cases} 
\text{eng}'(k) - \tau, & \text{eng}'(k) > \tau \\
0, & \text{eng}'(k) \leq \tau 
\end{cases} \tag{15}
\]

Where \( \text{eng}(k) \) denotes the energy value of topic \( k \) in slot \( t \). \( \tau \) denotes the decreasing ratio which is determined by the mean value of topic lifespan from training dataset. After the step of energy decay, the whole energy refreshing procedure is performed. The topic clusters with the positive energy value are considered as active topics, while
The topic clusters with non-positive energy value are transformed into dead clusters. In our algorithm, topic set composed of all active topics is defined as candidate topic set. In the following similarity calculation procedure in clustering, the calculation is constrained to the candidate topic set instead of all topics ever existed, which greatly improves the time efficiency of the algorithm.

By introducing candidate topic set into Single Pass Clustering, we propose our clustering algorithm as shown in Fig. 6.

The time complexity of candidate topic set selection procedure is 2n. Because the quantity of active topic clusters is significantly diminished, the time complexity of new algorithm is obviously less than the original algorithm while keeping a similar accuracy. The experiments in Sect. 4 show this point.

4. Experiments

4.1 Experimental Design

Though our proposed OTD method can be implemented in any online BBS environment, in order to evaluate the performance of online topic detection method for BBS conveniently, a pre-processed offline dataset is adopted in our experiments (described in Sect. 4.2), which can be used to simulate an online BBS environment to a large extent. Three experiments are designed as follows.

The first experiment is performance evaluation in accuracy. CDET [20] is used to measure the accuracy of the clustering algorithms and its calculation equation is presented later. This experiment includes two parts. The first part evaluates the strategies for representative post selection during the text preprocessing phase including simple representative post selection, keyword extraction and the Markov chain model. As we mentioned above, the keyword extraction strategy is proposed by N. Wanas et al [14] and the Markov chain model strategy is proposed by our work. The second part evaluates the strategies for candidate topic set generation during the topic clustering phase, including simple candidate strategy and Aging Theory. The Aging Theory strategy is used in both the work by D. Zheng and F. Li [13] and our work. By comparing the accuracy of the results obtained by the methods with different strategies in each phase independently, we evaluate our clustering algorithm conveniently.

The second experiment is performance evaluation in time complexity, which aims to compare the time efficiency of topic clustering with different candidate topic set generation strategies on relatively large data set. As the same as in the first experiment, the candidate topic set generation strategies include simple candidate strategy and Aging Theory, which is used in both the work by D. Zheng and F. Li [13] and our work. We execute Single Pass Clustering, Single Pass Clustering with simple candidate strategy and Single Pass Clustering with Aging Theory strategy on the same dataset respectively. And by comparing the time efficiency of the results obtained by the methods, we evaluate the time efficiency of our clustering algorithm conveniently.

The third experiment explores the influences of candidate topic set generation strategies on detection accuracy. In the topic clustering phase, we use Aging Theory to generate candidate topic set to improve the time efficiency. But it is necessary to explore how the candidate topic set generation strategies affect the detection accuracy. So we compare the results between Single Pass Clustering and Single Pass Clustering with simple candidate strategy or Aging Theory strategy. F1 is used to measure the similarity between two topic clustering results, its calculation equation is presented later.

We implemented our proposed method and the baseline algorithms in C++ programming language and executed these programs on the same computer. The computer is a ThinkPad T420 Laptop with Intel i5 2520M/2.5 GHz CPU and 4 GB memory.

4.2 Dataset and Parameter Training

As there is no public dataset available for BBS OTD, we collected one dataset from a real BBS called Newsmsn [8]. Newsmsn is a famous BBS forum in China with about 30,000 online members, more than 10 discussion boards and about 25,000 posts every day. The users of the Newsmsn forum are mainly university students and teaching or research staff.

The summary of the dataset is shown in Table 1.

Some of the post files in the dataset are of low quality or off-topic. As mentioned in Sect. 3, we pick out the high-quality posts inside each thread by representative post selection procedure based on Markov chain model.

The chosen posts are semi-structured html pages, which are always mixed with some noisy information. So text pre-processing is very critical for the future process-
The accuracy of OTD is measured by minimum normalized performance evaluation in accuracy following experiments.

\[ \alpha \text{ precision rate of 87.4}. \] Therefore, the procedure successfully classified the high-quality posts with precision rate of 87.4%.

Table 2 The percentage of high-quality posts identified by the representative post selection procedure with different parameter settings.

| Parameter | 0.2 | 0.4 | 0.6 | 0.8 | 0.82 | 0.84 | 0.86 | 0.9 |
|-----------|-----|-----|-----|-----|------|------|------|-----|
| S         | 587 | 962 | 766 | 1021 | 1109 | 1311 | 1276 | 855 |
| S(total%)  | 39.1 | 66.1 | 51.1 | 68.1 | 73.9 | 87.4 | 85.1 | 57.0 |

During the course of text combining, we extract the useful information including post headlines and post contents from the posts by the method based on RegEx.

After that step, the word segmentation and stop-words removal are executed on the extracted information of the posts. The ICTCLAS tool [21] developed by the Institute of Computing Technology of Chinese Academy of Science is used in word segmentation.

As described in Sect. 3, the Markov chain model is used to carry out the representative post selection. There is a parameter in the model called \( \alpha \), the reading transformation factor. The value of the parameter needs training.

We randomly selected 500 threads from the dataset and manually labeled the three most valuable posts for each thread. 1,500 high-quality posts are labeled in total. Then we execute our representative post selection procedure on these 500 threads to obtain the representative posts while adjusting the reading transformation factor \( \alpha \) which ranges from 0 to 1. We define \( S \) as the number of the posts which are in the 1,500 labeled posts. The part of the training results is shown in Table 2.

When \( \alpha \) is set to 0.84, the representative post selection procedure successfully classified the high-quality posts with precision rate of 87.4%. Therefore, \( \alpha \) is set to 0.84 in the following experiments.

### 4.3 Performance Evaluation in Accuracy

The accuracy of OTD is measured by minimum normalized \( C_{DET} \) [20], which is widely used in TDT evaluation. In this experiment, the minimum normalized \( C_{DET} \) of topic detection is chosen to represent the method’s accuracy.

\( C_{DET} \) indicates the combined cost of misses and false alarms. In general, the lower \( C_{DET} \) is, the more accurate OTD algorithm is. \( C_{DET} \) is obtained by solving the Eq. (16).

\[
C_{DET} = C_{Miss} \cdot P_{Miss} \cdot P_{target} + C_{FA} \cdot P_{FA} \cdot P_{non-target}
\]  

(16)

Where \( P_{miss} \) denotes the miss rate, \( P_{FA} \) denotes the false alarm rate. \( C_{Miss} \) and \( C_{FA} \) are the costs for misses and false alarms. \( P_{target} \) and \( P_{non-target} \) are prior probabilities for prediction \( (P_{target} + P_{non-target} = 1) \).

In this experiment, \( C_{Miss} \) is set to 1 and \( C_{FA} \) is set to 1. The normalized \( C_{DET} \) is obtained by solving the Eq. (17).

\[
(C_{DET})_{norm} = \min(C_{Miss} \cdot P_{target}, C_{FA} \cdot P_{non-target})
\]  

(17)

As we just mentioned, all the data for our experiments is raw data. It is very hard to divide almost 300,000 posts into different topics accurately. So this experiment about accuracy evaluation is implemented on a relatively small dataset as shown in Table 3.

About 70 topics are selected from the dataset manually. There are many threads in each topic. Because the threads without user reply are usually not representative enough, we filtered these threads with reply posts less than 7 out of the dataset, which is the average reply number of our dataset. At last, 2~10 threads are left in each topic.

As described in Sect. 3, we use TF-IDF model to generate thread vector. The term frequency of a word in a thread is described by the Eq. (6) in Sect. 3.

There is a parameter \( r \) in this equation which needs training. We perform the training program on the dataset as shown in Table 3. Our topic clustering algorithm is performed on different groups of thread vectors generated from the testing dataset by adjusting the parameter \( r \). The relation between the parameter \( r \) and the detection accuracy is shown in Fig. 7.

When the value of parameter \( r \) is between 4 and 6, the \( (C_{DET})_{norm} \) of the results is the lowest, which means that the detection accuracy is the highest. So we set the parameter \( r \) to 5 in the rest of the experiments.

As mentioned in Sect. 4.1, this experiment includes two parts. In the first part, we compare the strategies for representative post selection, including simple representative post selection, keyword extraction and the Markov chain model.

| Dataset Name | TrainDS | TestDS |
|--------------|---------|--------|
| Time span    | July 18, 2010 – August 31, 2010 | August 1, 2010 – August 5, 2010 |
| #Posts       | 96864   | 32934  |
| #Threads     | 14061   | 4587   |

**Table 2** The percentage of high-quality posts identified by the representative post selection procedure with different parameter settings.

**Table 3** The datasets for performance evaluation in accuracy: Training dataset (TrainDS) is for training dictionary, document frequency, and etc., testing dataset (TestDS) is for testing accuracy of clustering.

**Fig. 7** The training results of parameter \( r \).
Table 4 The comparison among the different strategies for representative post selection.

|          | SPC | SPC with simple representative post selection strategy | SPC with keyword extraction strategy | SPC with the Markov chain strategy |
|----------|-----|--------------------------------------------------------|-------------------------------------|----------------------------------|
| $P_{Miss}$ | 0.080 | 0.069                                                  | 0.066                               | 0.066                             |
| $P_{FA}$  | 0.139 | 0.135                                                  | 0.135                               | 0.128                             |
| $(C_{DET})_{norm}$ | 0.566 | 0.538                                                  | 0.534                               | 0.514                             |
| $TC(h)$  | 3.988 | 3.772                                                  | 3.763                               | 3.763                             |

Table 5 The comparison among the different strategies for topic clustering.

|          | SPC | SPC with simple candidate strategy | SPC with Aging Theory strategy |
|----------|-----|-----------------------------------|--------------------------------|
| $P_{Miss}$ | 0.080 | 0.091                            | 0.080                           |
| $P_{FA}$  | 0.139 | 0.146                            | 0.142                           |
| $(C_{DET})_{norm}$ | 0.566 | 0.593                            | 0.575                           |
| $TC(h)$  | 3.988 | 2.986                             | 2.525                           |

The simple representative post selection strategy removes the posts with less than 5 words, ranks the posts by the number of reply posts and selects top 7 posts as the representative posts for the thread. The keyword extraction strategy was proposed by N. Wanas et al. [14], its main idea is to assess the relevance between the posts and the thread based on auto keyword extraction. As described in Sect. 3.2, the Markov chain model strategy is proposed by us. The strategies for topic clustering include simple candidate strategy and Aging Theory strategy. The simple candidate strategy simply filters the clusters which don’t have new posts in 3 days. And the Aging Theory strategy, which is also used by D. Zheng and F. Li [13], is described in Sect. 3.4. Here, the full decay period of a topic is set to 25,000 posts, which means that the energy value of a topic will decrease from 1 to 0 if none of the 25,000 new posts relates to the topic. 25,000 is the approximate number of 3 days’ posts collected in 2010 (as shown in Table 6) are used in this experiment. As described in Sect. 4.3, the simple candidate strategy simply filters the clusters which don’t have the new posts in 3 days. And as to Aging Theory strategy, the full decay period of a topic is set to 25,000 posts.

The comparisons among the different strategies for representative post selection is shown in Table 4, and the comparison among the different strategies for topic clustering is shown in Table 5.

Table 4 shows that the method with any representative post selection strategy can achieve better accuracy by decreasing $P_{Miss}$, $P_{FA}$ and $(C_{DET})_{norm}$. Simple representative post selection strategy and keyword extraction strategy obtain close results, and the Markov chain model strategy is the best strategy which has the minimal $P_{FA}$ and $(C_{DET})_{norm}$, and $P_{Miss}$ and $TC(h)$ also don’t exceed other strategies. In addition, $TC(h)$ decreases a little because threads are simplified.

Table 5 shows that Simple Pass Clustering with any candidate topic set generation strategy can make the accuracy worse with a little larger $P_{Miss}$, $P_{FA}$ and $(C_{DET})_{norm}$. But the $TC(h)$ of Single Pass Clustering with candidate topic set generation strategy is less than Single Pass Clustering. The Aging Theory strategy behaves better than simple candidate strategy in both accuracy and efficiency.

From this experiment, we can see that representative post selection strategies can significantly improve the accuracy of OTD on BBS. And the Markov chain model strategy used in our clustering algorithm is the best strategy among three different strategies. The candidate topic set generation strategies for topic clustering decrease the accuracy in a totally acceptable degree, but significantly improves the efficiency of clustering. And Aging Theory strategy behaves better than simple candidate strategy in both accuracy and efficiency.

4.4 Performance Evaluation in Time Complexity

As mentioned above, the candidate topic set generation strategy based on simple candidate strategy or Aging Theory greatly improves the efficiency of topic clustering. In this experiment, we test the time complexity of clustering algorithms with different candidate topic set generation strategies on relatively large datasets. Different sizes of datasets collected in 2010 (as shown in Table 6) are used in this experiment.

As described in Sect. 4.3, the simple candidate strategy simply filters the clusters which don’t have the new posts in 3 days. And as to Aging Theory strategy, the full decay period of a topic is set to 25,000 posts.

The relations between the time cost and number of threads with Single Pass Clustering, Single Pass Clustering with simple candidate strategy and Single Pass Clustering with Aging Theory strategy are shown in Fig. 8.

Figure 8 shows that the time complexity of Single Pass Clustering grows considerably when large dataset is given, which verifies our analysis in the former section. Single Pass Clustering with any candidate topic set generation strategy achieves better efficiency. Our topic clustering algorithm with candidate topic set generation strategy based on Aging Theory outperforms the other two clustering algorithms and can obviously improve the efficiency by achieving nearly linear complexity, especially for large dataset.
4.5 Influences of Candidate Topic Set Generation Strategies on Detection Accuracy

In Sect. 4.3, we evaluate the accuracy of our clustering method on a relatively small dataset, because it is difficult and time consuming to label large datasets manually. The experimental results show that two candidate topic set generation strategies, simple candidate strategy and Aging Theory strategy, have some bad influences on detection accuracy.

In order to measure the influences of different candidate topic set generation strategies on detection accuracy, we compare the results between Single Pass Clustering and Single Pass Clustering with simple candidate strategy or Aging Theory strategy. The more similar the results of two algorithms are, the less the influence of the candidate topic set generation strategy is.

The datasets in this experiment include 6 datasets from TestDS1 to TestDS6 in Table 6. Through the experiment in Sect. 4.4, we get the clustering results of these 6 datasets using Single Pass Clustering, Single Pass Clustering with simple candidate strategy and our topic clustering algorithm with Aging Theory strategy. The clustering results are shown in Table 7.

From Table 7, we can see the difference among the results of three clustering algorithms. Generally, the cluster number of Single Pass Clustering is much smaller than that of the other two clustering algorithms. That is because Single Pass Clustering treats all threads equally while Single Pass Clustering with candidate topic set generation strategy pays more attention to the active threads. Finally, the clustering algorithms with two different candidate topic set generation strategies achieve similar results.

Because it is infeasible to label the topics for all 30,717 threads manually, the similarity between two clustering results is calculated as the approximate detection accuracy.

In our experiment, $F_\beta$ is used to measure the similarity between the two topic clustering results, which can be obtained by solving the following Eq. (18).

$$F_\beta = \frac{(\beta^2 + 1) \cdot p \cdot r}{\beta^2 \cdot p + r}$$  \hspace{1cm} (18)

Where $p$ denotes the precision ratio, $r$ denotes the recall ratio, $\beta$ denotes the weight adjustment between precision ratio and recall ratio. In this experiment, $\beta$ is set to 1, so $F_1$ is used to represent the similarity between the two clustering results, as shown in Eq. (19).

$$F_1 = \frac{2 \cdot p \cdot r}{p + r}$$  \hspace{1cm} (19)

The value of $F_1$ is obtained on different datasets between the results of Single Pass Clustering, and Single Pass Clustering with simple candidate strategy or Single Pass Clustering with Aging Theory strategy. $F_1$ values of Single Pass Clustering and Single Pass Clustering with simple candidate strategy or Aging Theory strategy in different dataset size are shown in Fig. 9.

In Fig. 9, with the growth of the number of threads, the $F_1$ values of Single Pass Clustering and Single Pass Clustering with simple candidate strategy, Single Pass Clustering and Single Pass Clustering with Aging Theory strategy

| Dataset name | Cluster number of SPC | Cluster number of SPC with simple candidate strategy | Cluster number of SPC with Aging Theory strategy |
|--------------|-----------------------|--------------------------------------------------|-----------------------------------------------|
| TestDS1      | 1363                  | 1943                                             | 1829                                          |
| TestDS2      | 2100                  | 3834                                             | 3848                                          |
| TestDS3      | 2624                  | 6372                                             | 5768                                          |
| TestDS4      | 3462                  | 6903                                             | 7427                                          |
| TestDS5      | 5490                  | 8987                                             | 9018                                          |
| TestDS6      | 6987                  | 11323                                            | 10599                                         |
will decrease. The $F_1$ value of Single Pass Clustering and Single Pass Clustering with Aging Theory strategy is larger than that of Single Pass Clustering and Single Pass Clustering with simple candidate strategy, which indicates that the candidate topic set generation strategy based on Aging Theory used in our clustering algorithm has less influence on detection accuracy.

From the three experiments above, we can find that both accuracy and time efficiency should be considered when we choose clustering algorithms.

5. Discussion on Aging Characteristic of Topics

As mentioned in Sect. 1, one of the BBS characteristics is that the generation and aging of the topics on BBS are very fast. This characteristic is a very important inspiration to our topic clustering algorithm.

The candidate topic set generation strategy in our method is based on the aging characteristic of the topic. This experiment shows that the generation and aging of the topics on BBS are very fast.

The training dataset and the testing dataset are shown in Table 8.

After the generation of the clustering result using the testing dataset, we can find the 3 hottest topics in August 2010. These topics are Guo Degang event, tourist bus hijacked in Philippines and shopping discussion. Then we count the posts under each topic every day. The trends are shown in Fig. 10.

The topic trends in August 2010 show that there is a life cycle of each topic. The topic can be active only for a couple of days (as shown in Fig. 10, the hottest topic continues for 8 days). So the generation and aging of the topics on BBS are very fast.

6. Conclusions and Future Work

Because of the distinct characteristics of text, the existing technologies of OTD for news web pages are not suitable for BBS. We summarized the characteristics of BBS including organizational flexibility of BBS texts, high data volume and aging characteristic of BBS topics. Based on these distinct characteristics of BBS, a novel method of OTD for BBS is proposed, firstly a representative post selection strategy based on post quality ranking is used to choose 7 representative posts from a thread, and then incremental TF-IDF model is used to transform representative texts into thread vector, finally an efficient topic clustering algorithm with candidate topic set generation strategy is used to cluster these vectors. At the same time, aging theory is used to minimize the size of candidate cluster set. The experimental results show that our method improves the performance of OTD for BBS environment on both accuracy and time complexity significantly.

Our representative post selection method considers the topological relationship between posts without considering the actual content of the posts. In the future, we plan to incorporate the content characteristics, which are expected to improve the accuracy of representative post selection further. We also plan to apply our proposed method into microblogs, a trendy social networking service in recent years.

Acknowledgments

Our work is supported in part by the National Natural Science Foundation of China under Grant No.61103131 and the National Key Technology R&D Program of China under Contract No. 2012BAH23B03.

References

[1] http://www.cnbeta.com/articles/208347.htm
[2] http://www.cnbeta.com/articles/218763.htm
[3] http://digi.it.sohu.com/20120412/n340366938.shtml
[4] http://data.cnzz.com
[5] Y. Yang, T. Pierce, and J. Carbonell, “A study of retrospective and on-line event detection,” Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 98), pp.28–36, New York, USA, Aug. 1998.
[6] J. Allan, R. Papka, and V. Lavrenko, “On-line new event detection and tracking,” Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 98), pp.37–45, New York, USA, Aug. 1998.
[7] http://ldc.upenn.edu
[8] http://www.newsml.net
[9] J. Fiscus, G. Doddington, J. Garofolo, and A. Martin, “Nist’s 1998 topic detection and tracking evaluation (TDT2),” Proc. the DARPA Broadcast News Transcription and Understanding Workshop, pp.19–26, San Francisco, CA, 1999.
[10] Q. He, K. Chang, E. Lim, and A. Banerjee, “Keep it simple with time: A reexamination of probabilistic topic detection models.”
IEEE Trans. Pattern Anal. Mach. Intell., vol.32, pp.1795–1808, Oct. 2010.

[11] M. Zhu, W. Hu, and O. Wu, “Topic detection and tracking for threaded discussion communities,” Proc. ACM International Conference on Web Intelligence (WI 08), pp.77–83, Dec. 2008.

[12] J. Kim, K. Candan, and M. Dönderler, “Topic segmentation of message hierarchies for indexing and navigation support,” Proc. World Wide Web Conference Series (WWW 05), pp.322–331, May 2005.

[13] D. Zheng and F. Li, “Hot topic detection on BBS using aging theory,” Proc. Web Information Systems and Mining (WISM 09), pp.129–138, Nov. 2009.

[14] N. Wanas, A. Magdy, and H. Ashour, “Using automatic keyword extraction to detect off-topic posts in online discussion boards,” Proc. Content Analysis for the WEB 2.0 (CAW 2.0 09), April 2009.

[15] T. Brants and F. Chen, “A system for new event detection,” Proc. 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 03), pp.330–337, July 2003.

[16] J. Schultz and M. Liberman, “Topic detection and tracking using IDF-weighted cosine coefficient,” Proc. DAPPA Broadcast News Workshop, pp.189–192, July 1999.

[17] T. He, G. Qu, S. Li, X. Tu, Y. Zhang, and H. Ren, “Semi-automatic hot event detection,” Proc. Advanced Data Mining and Applications (ADMA 06), pp.1008–1016, Aug. 2006.

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

[19] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank citation ranking: Bringing order to the Web,” Google Company, 1999.

[20] C.J. van Rijsbergen, Information Retrieval, 2nd ed., Butterworths, Massachusetts, 1979.

[21] http://www.ictclas.org

Appendix

Portions of the work reported here were previously presented in a full paper by Yan Zhao and Jungang Xu on page 453 to 457 in the proceedings of 2011 IEEE 3rd International Conference on Communication Software and Networks (ICCSN 2011). The extensions to the conference version include: (i) We introduce Aging Theory into our topic clustering algorithm to generate candidate topic set; (ii) the size of the dataset used in this paper is quadrupled to the dataset in that paper; (iii) separately test the influences that representative post selection strategy and candidate topic set generation strategy bring on the accuracy of the OTD method; (iv) about 6 datasets of different sizes are used in the evaluation of time complexity of the related algorithms; (v) the evaluation of the influence on detection accuracy from the candidate topic set generation strategy is carried out on these 6 datasets in different size; (vi) a discussion on the generation and aging characteristic of BBS topics is presented.