Chinese Microblog Topic Detection Based on the Latent Semantic Analysis and Structural Property

Xia Yan
Shenzhen Institute of Information Technology, Shenzhen, China, 518172
Email: yanx@sziit.com.cn

Hua Zhao
College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao, China, 266590
Email: doctorhuazhao@yahoo.com.cn

***Abstract*—traditional topic detection method can not be applied to the microblog topic detection directly, because the microblog text is a kind of the short, fractional and grass-roots text. In order to detect the hot topic in the microblog text effectively, we propose a microblog topic detection method based on the combination of the latent semantic analysis and the structural property. According to the dialogic property of the microblog, our proposed method firstly creates semantic space based on the replies to the thread, with the aim to solve the data sparseness problem; secondly, create the microblog model based on the latent semantic analysis; finally, propose a semantic computation method combined with the time information. We then adopt the agglomerative hierarchical clustering method as the microblog topic detection method. Experimental results show that our proposed methods improve the performances of the microblog topic detection greatly.**

***Index Terms*— microblog, topic detection, semantic space, latent semantic analysis**

I. INTRODUCTION

Microblog is an information publication, spread and achievement platform based on the relationships between the users. Based on this platform, users can create personal community through web, wap and other manners, and public their information within 140 words. The spread speed of the topic in microblog is very rapid, because the microblog text is very short and on the other hand, the microblog platform has so many users, which makes microblog an important place for monitoring. So it is important to detect the user interested topic from the microblog text rapidly.

Topic detection is a task of the Topic Detection and Tracking (TDT), and is defined to be the task of automatically detecting new topics in the news stream and associating incoming stories with topics created so far [1]. Topic detection is essentially similar to the unsupervised clustering except that it is done incrementally, not globally [2]. The idea of topic detection is shown in Figure 1.

The traditional topic detection is defined to deal with the long news stories coming from the news Medias. Many researchers carried out researches into the traditional topic detection. But, microblog text is a kind of grass-roots text, and any registered user can public any information about any topics. Different user will have different wording style, so the traditional topic detection can not be applied to microblog topic detection directly.

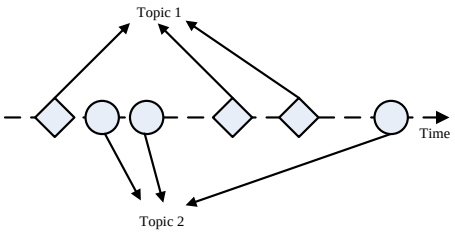

Figure 1 Main Idea of Topic Detection

In order to detect the hot topics in the microblog, we propose a microblog topic detection method based on the combination of latent semantic analysis and the structural property. According to the dialogic property of the microblog, our proposed method firstly creates semantic space based on the replies to the thread, with the aim to solve the data sparseness problem; secondly, create the microblog model based on the latent semantic analysis; finally, propose a semantic computation method combined with the time information. We then adopt the agglomerative hierarchical clustering method during microblog topic detection. Experimental results showed that our proposed methods improve the performances of the microblog topic detection greatly.

The structure of the paper is as follows. Section 2 gives a short overview of the best current approaches in the microblog topic detection research. Section 3 presents our system architecture used in the topic detection. Section 4 laid an emphasis on the microblog model based on the semantic space and latent semantic analysis. Section 5 covers the topic detection method combined with time information. Section 6 discusses the experimental results

doi:10.4304/jnw.8.4.917-923

and analysis. Section 7 gives the conclusions inferred from our work.

II. Related Work

A. Property Analysis of Microblog

Microblog is a new form of communication in which users can discuss their interested topics in short posts. Compared with the other communication platform, the microblog text has the following properties:

(1) Short. The length of the microblog text is limited to 140 words, which is much shorter than other network text, for example, news story, which will lead to the data sparseness problem.

(2) Grass-roots. Microblog platform allows any registered user to publish their opinions or any other interesting things. Different users have the different word-styling, and they will usually adopt different words to express the same idea. The grass-roots property makes the microblog topic detection more difficult.

(3) Political. Once a real event happens, there will have many discusses about this event on the microblog. Some other hot topics even launched in microblog. So it is important to monitor the microblog information to detect the hot topic.

There are also other properties, such as mass data, real-time updates, and so on. These properties make the traditional topic detection method can not be applied to microblog topic detection. But, on the other hand, the microblog structure gives us a new idea to improve the performance of the microblog topic detection. In this paper, we will make the best of these structural properties, and propose an effective microblog topic detection method which aims at the above three properties.

B. Related Work in Microblog Topic Detection

Many researchers have begun to carry out researches to the microblog topic detection, but most of these work are based on the English Twitter [3, 4], and the topic detection for Chinese microblog is just begin. The main technologies within the microblog topic detection include the microblog representation model and the topic detection method.

There are two main microblog representation models, which are Vector Space Model (VSM) based on TFIDF [5] and LDA model [6]. Ma Bin compared the performance of these two models in the microblog topic detection, and found that the performance of LDA model is better than that of VSM based on TFIDF [7, 8].

Some other researchers devote to resolve the data sparseness problem based on the expansion to the texts. For example, Bharati Sriram[9] expanded Twitter texts based on the user's profile; Ishikawa, S. [10]expanded the microblog texts based on Wikipedia, which has got better performance. More researchers expanded the microblog texts based on the semantic dictionary [11]. These expansion methods based on the external resources can improve the data sparseness problem to a certain extent, but they sometimes will introduce noise.

The microblog topic detection algorithm is still based on the clustering method, for example Single-Pass [12, 13], K-Means, HMM [14] and Bayesean clustering [15], and so on. Many users explored to combine the structural property on the basis of the traditional topic detection algorithm. Based on the dialogic property of microblog, Ma Bin proposed a thread-based two-stage clustering method [7]. Rui Long proposed a microblog topic detection method based on the word co-occurrence graph [16], which firstly creates the word co-occurrence graph, and then regard the disconnected cluster as a news topic. The research of C Akcora showed that the number of the emotional words will increase when a hot event happened [17]. Based on this foundation, Lin Hongfei proposed an emotional language model, which realized the hot event detection based on the analysis of the difference between emotional language models in adjacent periods [18].

C. Introduction to the Latent Semantic Analysis

Latent Semantic Analysis (LSA) is also called Latent Semantic Indexing (LSI), which is a mathematics technique for creating vector-based representations of texts, with the aim to map the text model based on the high-dimensional Vector Space Model (VSM) to the low-dimensional latent semantic space. It is a common-used technique in information retrieval, and some other areas [19], and performs well.

The process of the LSA is firstly to analyze the massive texts, create a word-text matrix to describe the occurrence of the words in the texts, and then extract the latent semantic structure between the words to represent the words and the texts.

The idea of LSA thinks that there are some kinds of relationship between the words and the texts. The texts and the words can create certain semantic structure based on their relationship. Then, we can compute and handle the structure using the mathematical principles and methods, and hold the main relationship between the texts and the words. If we use the matrix $A_{m\times n}$ to represent the words-texts co-occurrence matrix, $A_{m\times n}=[\,a_{ij}\,]_{m\times n}$, where m is the number of the different words in these texts, n is the number of the texts, and a_{ij} is the weight of the i^{th} words in the j^{th} text, with the typical weighting method is TF-IDF. $A_{m\times n}$ will be a very sparse matrix.

And then carry out Singular Value Decomposition (SVD) to $A_{m\times n}$, and achieve a new matrix A_k to act as the approximate matrix of $A_{m\times n}$, where $k<<min(m, n)$. The decomposition process is as follows [20]:

(1) Firstly, create the term-document co-occurrence matrix A, $A=\begin{bmatrix} a_{1,1}\cdots a_{1,n} \\ \vdots \ddots \vdots \\ a_{m,1}\cdots a_{m,n} \end{bmatrix}$, and each row $t_i^T=\begin{bmatrix} a_{1,1}\cdots a_{1,n} \end{bmatrix}$ in this matrix will be a vector

corresponding to a term, and each column $d_j = \begin{bmatrix} a_{1,j} \\ \vdots \\ a_{m,j} \end{bmatrix}$ will be a vector corresponding to a document.

(2) Now, we assume there are a decomposition of A : $A = U\Sigma V^T$, where U and V are orthogonal matrices, Σ is a diagonal matrix.

(3) Then we will have: $AA^T = U\Sigma\Sigma^{\mathrm{T}}U^T$ and $A^T A = V\Sigma^{\mathrm{T}}\Sigma V^T$. Since $\Sigma\Sigma^{\mathrm{T}}$ and $\Sigma^{\mathrm{T}}\Sigma$ are diagonal, so U and V must contain the eigenvectors of AA^T and $A^T A$, respectively.

$$A = (\hat{t}_i^T) \to \left[\begin{bmatrix} & \\ u_1 & \end{bmatrix} \cdots \begin{bmatrix} & \\ u_l & \end{bmatrix}\right] \cdot \begin{bmatrix} \sigma_1 \cdots 0 \\ \vdots \ddots \vdots \\ 0 \cdots \sigma_l \end{bmatrix} \cdot \begin{bmatrix} [v_1] \\ \\ [v_l] \end{bmatrix} \quad (1)$$

The values $\sigma_1 \cdots \sigma_l$ are called the singular values, and $u_1 \cdots u_l$ and $v_1 \cdots v_l$ are the left and right singular vectors.

(4) The related researches show that when we select k as the largest singular values, and their corresponding singular vectors from U and V , we can get the rank k approximation to A with the smallest errors. But more importantly, we can now treat the term and document vectors as a "semantic space". The vector $\hat{t}_i$ then has k entries mapping it to a much lower dimensional space dimensions. We can write the approximation as: $A_k = U_k \Sigma_\kappa V_k^T$.

III. System Architecture

Our system architecture used in our topic detection system is shown in Figure 2.

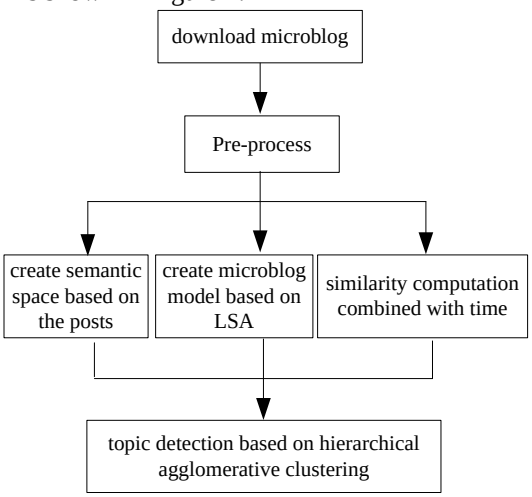

Figure 2 System Architecture

In our following experiments, in order to evaluate the performances, we firstly download the microblog from Sina manually. Micro-blog text is a kind of user generated content (UGC), a kind of heavy grass-roots text, and its content usually include much noise data, so before the topic detection, we must pre-process the downloaded microblog texts, which is very important to the success of our detection system. Now, our adopted pre-processes are as follows:

1. Microblog texts cleaning: There are many noise data in the microblog text, for example, user account, emoticons and URL, which will make a bad influence on the accuracy of the similarity computation between two microblog texts. In order to get an effective topic detection system, microblog text cleaning firstly remove these noises based on the following methods:

(1) Delete URL. There are many URLs in the microblog, which is mainly because that the length of a micro-blog text is limited to be no more than 140 words, in order to give detailed introduction to a certain topic, users usually add a URL in its current microblog text, where will include the detail information. Some researchers make an effort to explore the usage of these URLs. Because the URLs are usually have the unified format, so we choose to delete them using regular expression.

(2) Remove user account. In the Sina Micro-blog, there are many user accounts, which are usually used in two cases, and we adopt different operations to these two cases. For the first case, when the current micro-blog A is the reply to the micro-blog B, the user account is usually used as "//@user account" to cite the content of B. We then will delete the string "//@user account" completely for this case. The second case is when a user writes his/her micro-blog, he (she) sometimes mentions another microblog user. The user account will be used as "@user account" in this case. We will only delete the symbol "@" for this case because the user account is the actual content of the microblog.

(3) Delete emoticons. Emoticons are common in the Internet, also in the microblog text. Users usually use these emoticons to express their ideas about a certain topic. These emoticons are very useful for the sentiment analysis, but they are of lesser significance in keyword extraction. The emoticons are usually converted into strings with unified format, that is, the strings are enclosed by bracket, for example [sunlight], [doubt] and so on. According to the above foundlings, we delete the emoticons based on the predefined rules.

2. Because we mainly detect the hot topic from Chinese microblog, so we secondly carry out Chinese Word Segment (CWS) and part of speech (POS) tagging. In our experiments, we adopt the ICTCLAS2011 (http://ictclas.org/ictclas_download.aspx) to segment and tag, which is shared by Chinese Academy of Sciences.

3. Beside the above pre-processes, we also carry out some other pre-processes, which mainly include removing stop word. Because the words with certain POS can't possibly be keyword, for example, preposition, conjunction, and so on.

Ⅳ. Microblog Modeling Method Based on the Posts and Latent Semantic Analysis

Before we create the model of the microblog, we apply the pre-processing to the pieces of the mircoblog and

their replies, which include Chinese word segmentation, removing the stop words, the emoticons and URL.

A. Creating Microblog Semantic Space Based on the Posts

The short text property of microblog leads to the heavy data sparseness problem, which is a very difficult problem in the microblog topic detection research. So, in order to solve this problem, we propose to create the miciroblog semantic space to expand the content of the microblog.

The typical microblog structure is as follows:

Figure 3 Microblog Structure

From the above figure, we can see that when a user publish a thread, some others users will give their opinions when they interested in this topic. In order to protect the privacy of the users, we cover the user ID by the blue rectangle in Figure 2. The threads and the posts are all called microblog texts in this paper.

Normally, a topic will become a hot topic when many users are interested in this topic. There are many replies to this topic accordingly, and all these replies are usually about the same topic with the thread of the microblog. These replies can resolve the data sparseness problem effectively, so we propose to create the microblog semantic space, which consists of the thread of the microblog and all the replies to the thread. The formal definition is as follows:

$$SemanticSpace(T) = \{T, r1, r2, ..., rn\} \quad (2)$$

where T means the thread of the microblog, and the $ri(1 \le i \le n)$ is the reply to the thread.

In order to increase the effectiveness of the expansion, the replies will firstly be filtered. That is the replies after the filtering can be used to expand the thread of microblog. All the replies have been pre-processed; we then remove the emotional words according to the emotional words list of HowNet. If the length of the reply (after removing the emotional words) is less than 4, the reply will be deleted.

B. Microblog Modeling Methods Based on Latent Semantic Analysis

After the works in the section, every thread of the microblog corresponds to a semantic space. Now, we adopt the latent semantic analysis to create the model of the semantic space.

(1) Create the word-semantic space co-occurrence matrix, $A_{m \times n}$. After the pre-process, we get 19.872 semantic spaces which include 82,365 words, that is m=82365 and n=19872, so we have the $A_{82365 \times 19872}$ matrix. a_{ij} is computed TFIDF, a classical and traditional method, which is computed by the follows:

$$a_{ij} = \frac{tf_{ij} \times \log_2(N/n_i + 0.01)}{\sqrt{\sum_{j=1}^{N}(tf_{ij} \times \log_2(N/n_i + 0.01))^2}} \quad (3)$$

where tf_{ij} is the term frequency of the i^{th} word in the j^{th} semantic space. N is the total number of the semantic spaces, n_i is the total number of the semantic spaces which include the i^{th} word.

(2) After we get the matrix $A_{82365 \times 19872}$, we need to do Singular Value Decomposition to $A_{82365 \times 19872}$. In our experiments, we adopt SVDLIBC to do SVD, which is developed by Massachusetts Institute of Technology. After the decomposition, we will get approximate matrix A_k, and at the same time, we will have the word vector U_k and the semantic space vector V_k.

V. Microblog Topic Detection Method Combined Time Information

Based the above section, we get the model of the semantic space; the next work is to realize microblog topic detection based on the agglomerative hierarchical clustering method. One critical step of the clustering is to compute the similarity between the two semantic spaces. Based on the Cosine, we propose a new similarity computation method combined with the time information.

The spread process of the topic has the property of the time concentration, which means that the stories related to the topic will be less and less. That is to say that a topic will not be concerned after a period of time. Based on this, we propose a new similarity computation method combined the time information, as follows:

$$Sim(S1, S2) = \frac{1}{|T(S1) - T(S2)|} \times Cosine(S1, S2) \quad (4)$$

$$Cosine(S1, S2) = \frac{\sum_{k=1}^{n} w_{s1\_k} \times w_{s2\_k}}{\sqrt{\sum_{k=1}^{n} w_{s1\_k}^{\ 2} \times \sum_{k=1}^{n} w_{s2\_k}^{\ 2}}} \quad (5)$$

where *S1* and *S2* represent two semantic spaces, and they are represented by two vectors based on the latent

semantic analysis: S1= {s1_1, s1_2… s1_n} and S2= {s2_1, s2_2… s2_n}. T(S1) and T(S2) represent the publication times (in minute) of the threads of the semantic spaces, respectively. From the formula, we can see that the bigger the time difference between the two threads, the smaller the similarity between the two semantic spaces.

We then adopt the agglomerative hierarchical clustering method to realize the microblog topic detection, and the end condition of the clustering method is the similarity between any clusters is not larger than the threshold θ. We do several experiments when is set to the different values, and we find that the systems perform best when is equal to 0.3.

VI. Experiments and Results Analysis

A. Corpus

Now, the microblog topic detection research is just beginning, so there is no public microblog corpus. In order to evaluate the methods proposed in our paper, we collect 15,850 pieces of microblog manually from Sina, which include 11 topics. The topics in the corpus include the "limited purchasing of milk", "divorce for the house" and "two sessions", and so on. In order to simulate the real environment, we arrange these pieces of microblog in chronological order, and save the corpus with the following uniform format:

```
<MICROBLOG>
  <ID>00001</ID>
  <USER>春意来了</USER>
  </TIME>2013-2-10 10:09</TIME>
  <TEXT>天气越来越暖和了, 真好</TEXT>
  <RT1>是时候考虑春游了</RT1>
  <RT2>喜欢春天</RT2>
</MICROBLOG>
```

where <ID> represents the number of the microblog, <USER> represents the user name who write this microblog, and <TIME> represents the microblog publication time, <RT> represent the reply to this miciro, if there are 10 replies, the serial number will be from 1 to 10, that is from <RT1> to <RT10>.

B. Evaluation Metrics

We adopt the evaluation metrics used in TDT to evaluate our systems, which include the miss rate, the false rate and the normalized cost. If the detection results of the i^{th} topic are listed in Table 1, then the miss rate ($Miss(i)$) and the false rate ($Fallout(i)$) of the i^{th} topic can be computed using (6) and (7), respectively.

TABLE Ⅰ.
MEANINGS OF THE EVALUATION PARAMETERS

	Related Microblog	Not Related Microblog
Detected Microblog	a	b
Not Detected Microblog	c	d

$$\mathrm{M}iss(i) = \frac{c}{a+c} \tag{6}$$

$$Fallout(i) = \frac{b}{b+d} \tag{7}$$

Once we obtain the performance (miss rate and fall rate) for every topic, we can achieve the system performance by averaging these metrics. The average miss rate (P_{miss}), average false rate (P_{Fall}) and the normalized cost ($(C_{Det})_{Norm}$) can be computed by (7), (8) and (9), respectively.

$$P_{\mathrm{Miss}} = \frac{1}{n}\sum_{i=1}^{n} Miss(i) \tag{8}$$

$$P_{Fall} = \frac{1}{n}\sum_{i=1}^{n} Fallout(i) \tag{9}$$

$$(C_{Det})_{Norm} = \frac{C_{Miss} \times P_{Miss} \times P_{t\mathrm{ar}} + C_{FA} \times P_{Fall} \times P_{\neg t\mathrm{ar}}}{\min(C_{Miss} \times P_{t\mathrm{ar}}, C_{FA} \times P_{\neg t\mathrm{ar}})} \tag{10}$$

where $P_{t\mathrm{ar}}$ is the probability of seeing a new story in the stream; C_{Miss} is the cost of missing a new story; $P_{\neg t\mathrm{ar}}$ is the probability of seeing an old story, $P_{\neg t\mathrm{ar}}$ =1- $P_{t\mathrm{ar}}$; C_{FA} is the cost of a false alarm; The values of C_{Miss} , C_{FA} , and $P_{t\mathrm{ar}}$ are usually predefined according to the application, in our experiments, C_{Miss} , C_{FA}, and $P_{t\mathrm{ar}}$ are set to 1.0, 0.1, and 0.02, respectively. The smaller the normalized cost $(C_{Det})_{Norm}$, the better the performance of the systems.

C. Experiments and Results

In order to verify the validity of these methods proposed in this paper, we design the following six systems, which are noted as BasicSys, ModiSys1, ModiSys2, ModiSys3, ModiSys4, and ModiSys5. The settings of these systems are shown in TABLE II, where different system use different techniques.

TABLE II. EXPERIMENTS SETTINGS

System Name	VSM	Cosine	Create Semantic Space	LSA	Similarity Combine Time
BasicSys	√	√			
ModiSys1	√	√	√		
ModiSys2	√	√		√	
ModiSys3	√				√
ModiSys4	√	√	√	√	
ModiSys5	√		√	√	√

The "*BasicSys*" is our baseline, and the systems from the "*ModiSys1*" to "*ModiSys5*" are our modified systems. All the six systems adopt the agglomerative hierarchical clustering method.

We firstly do some experiments under the different value of the similarity threshold θ for our baseline system. Figure 4 gives the comparisons between the system performances when the similarity threshold is set to different values.

From the results shown in Figure 4, we can see that the baseline system *BasicSys* achieve the best performance when the threshold is equal to 0.3, which is a relatively lower threshold. When we carry out deep analysis into the experimental results, we find that because microblog text is a heavy grass-roots text, many people will give very different discuss to a certain topic, the same words the users adopt will be very few, so the similarity

between two related microblog texts are relative smaller. So the similarity threshold will not very high.

So, in order to evaluate the performances of our proposed methods equally, we carry out all the other experiments (shown in TABLE II) when the similarity threshold is set to 0.3, and the experimental results (miss rate, fallout rate and normalized detection cost) are listed in TABLE III.

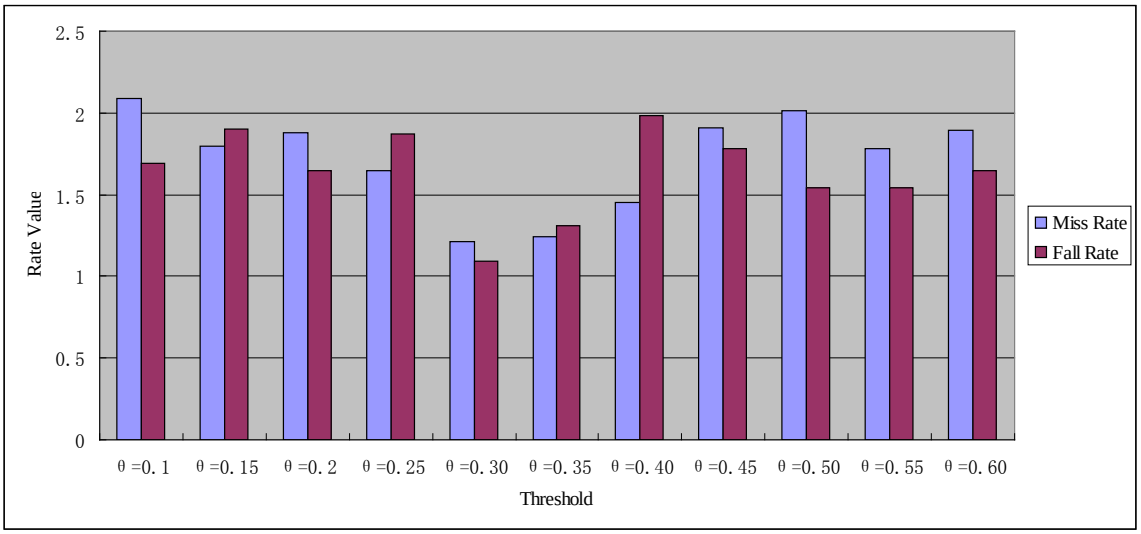

Figure 4 Comparisons of the Experimental Results

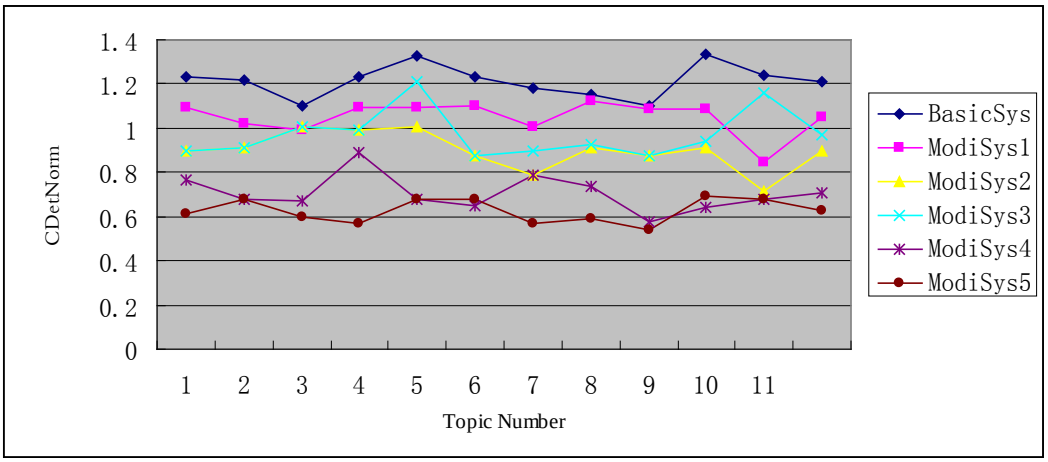

Figure 5 Comparisons of the Experimental Results

TABLE III.
EXPERIMENTAL RESULTS

System Name	P_{Miss}	P_{Fall}	$(C_{Det})_{Norm}$
BasicSys	0.3582	0.1706	1.2123
ModiSys1	0.2930	0.1543	1.0491
ModiSys2	0.2732	0.1278	0.8994
ModiSys3	0.3002	0.1372	0.9725
ModiSys4	0.2424	0.0943	0.7045
ModiSys5	0.2020	0.0865	0.6259

From the results shown in TABLE III, we can see that our proposed method improve the performance of the detection systems to different extent, which verify the effectiveness of our proposed methods. From the results, we can see that the false rate of the system decreases greatly when we adopt the time information when compute the similarity.

In order to show our results more clearly, we give the following experimental results comparison graph, shown in Figure5, which show the performance of these six systems over the 11 topics.From the comparisons shown in the Figure 5, we can see that *ModiSys5* get the best performances in all the six systems. That is to say that when we combine all the proposed method in one system, the system will perform best. So in the latter work, we can combine all the methods in a unique system.

VII. CONCLUSIONS

Topic Detection is a automatic technology to detect the hot topic in the news stories, which had aroused much attentions form the researchers, but the traditional topic detection method can be applied to the microblog texts directly.

Now, Microblog is one of the network Medias which develops rapidly and has widely influences on the people's life. In order to detect the hot topics in the microblog, we research deeply into the microblog topic

detection method, and propose a microblog topic detection method based on the latent semantic analysis and structural property. Experimental results show that LSA can resolve the wording diversity problem leaded by the grassroots property of microblog, the semantic space created based on the posts can resolve the data sparsity problem to some extent, and the similarity computation method combined with time can decrease the miss rate greatly. So, our proposed methods work successfully.

ACKNOWLEDGMENT

This work was supported by Sci & Tec Project of Shenzhen Institute of Information Technology (No.YB201016), the China Postdoctoral Science Foundation (2011M501155); NSFC under Grant (No.61170079 and No.61202152); the Special Fund for Fast Sharing of Science Paper in Net Era by CSTD (2012107).

REFERENCES

[1] K. Kamaldeep, V. Gupta, "A survey of topic tracking techniques", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 5, no. 2, 2012, pp. 383-392.

[2] D. Yongping, L. Jiangli, H. Ming, "Research on technique of the cyberspace public opinion detection and tracking", *2011 International Conference on Intelligent Computing and Information Science*, 2011, pp. 203-207.

[3] T. Sakaki, M. Okazaki, Y. Matauo, "Earthquake shakes twitter user: real-time event detection by social sensors", *Proceedings of the 19th International Conference on World Wide Web*, 2010, pp. 851-861.

[4] S. Phuvipadawat, T. Murata, "Breaking news detection and tracking in Twitter", *Proceedings of the 2010 International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'10)*, Toronto, Canada, Aug31-Sep 3, 2010, pp. 120-130.

[5] T. Wei, C Wei, M. Xiaofeng, "EDM: an efficient algorithm for event detection in microblogs", *Journal of Frontiers of Computer Science and Technology*, 2012, pp. 1-12.

[6] Z. Silong, L. Junyong, L. Yan, Y. Dong, Y. Tian, "Hotspots detection on microblog", *2012 Fourth International Conference on Multimedia Information Networking and Security*, 2012, pp. 922-925.

[7] M. Bin, H. Yu, L. Jianjiang, Y. Jianmin, Z. Qiaoming, "A thread-based two-stage clustering method of microblog topic detection", *Journal of Chinese Information Processing*, vol. 26, no. 6, 2012, pp. 121-128.

[8] Z. Jianfeng, X. Yunqing, M. Bin, Y. Jianmin, Y. Hong, "Thread cleaning and merging for microblog topic detection", *Proceedings of the 5th International Conference on Natural Language Processing*, 2011, pp. 589-597.

[9] B. Sharifi, M. A. Hutton, J. Kalita, "Summarizing micro-blogs with topic models", *Proceedings of NAACL-HLT 2010*, pp. 685-688.

[10] S. Ishikawa, Y. Arakawa, Tagashira S. Fukuda, "Hot topic detection in local areas using Twitter and Wikipedia", *ARCS Workshops (ARCS)*, 2012, pp. 1-5.

[11] L. Zitao, Y. Wenchao, C. Wei, "Short text feature selection for microblog mining", *The 4th International Conference on Computational Intelligence and Software Engineering*, Wuhan, China, 2010, pp. 1-4.

[12] B. Huang, Y. Yang, A. Mahmood, W. Hongjun, "Microblog topic detection based on LDA model and single-pass clustering", *Rough Sets and Current Trends in Computing Lecture Notes in Computer Science* ,vol. 7413, 2012, pp. 166-171.

[13] D. Yanyan, H. Yanxiang, Y. Tian, C. Qiang , L. Lu, "Microblog bursty topic detection based on user relationship", *2011 6th IEEE Joint International Conference on Information Technolog and Artificial Intelligence*, 2012, pp. 260-263.

[14] J. Hai-yan, W. Xing-ce, W. Zhong-ke, Z. Ming-Quan, W. Xue-Song, "Topic information collection based on the Hidden Markov Model", *Journal of Networks*, vol. 8, no.2, 2013, pp. 485-492.

[15] M. Mahbubur Rahman, "Unsupervised natural image segmentation using mean histogram features", *Journal of Multimedia*, vol. 7, no. 5, 2012, pp. 332-340.

[16] R. Long, W. Haofen, C. Yuqiang, J. Ou, and Y. Yong, "Towards effective event detection, tracking and summarization on microblog data", *Lecture Notes in Computer Science* Volume 6897, 2011, pp. 652-663.

[17] C. Akcora, M. Bayir, M. Demirbas, H. Ferhaosmanoglu, "Identifying breakpoints in public opinion", *Proceedings of KDD Workshop on Social Media Analytics*, 2010.

[18] Y. Liang, L. Yuan, L. Hongfei, "Micro-blog hot event detection based on emotion distribution", *Journal of Chinese Information Processing*. 2012, 26(1), pp. 84-91.

[19] S. Baoming, "A novel image correlation matching approach", *Journal of Multimedia*, vol. 5, no. 3, 2010, pp. 268-275.

[20] http://en.wikipedia.org/wiki/Latent_semantic_analysis

Xia Yan, born in JiLin, 1978. She received the bachelor degree in computer science and technology from Northeast Normal University, Chang Chun, China, in 2001, and received the master degree in computer science and technology from Harbin Institute of Technology, Harbin, China, in 2004.

She now is a lecture of Shenzhen Institute of Information Technology, Shenzhen, China. Her research interests include Question Answer, Topic Detection and Tracking, Natural Language Processing. She has published 5 papers related to Natural Language Processing. She has published 2 books, and now takes charge of a project.

Hua Zhao, born in Shandong, in 1980. She received the bachelor degree in computer science and technology from LiaoCheng University, China, in 2001, the master degree in computer science and technology from Harbin Institute of Technology (HIT), China, in 2003. She received the doctor degree in HIT in 2008.

She currently is a lecture in College of Information Science and Engineering, Shandong University of Science and Technology, China. Her research interests include Topic Detection and Tracking, Natural Language Processing, Machine Learning.