Popular Topic Detection in Chinese Micro-Blog Based on the Modified LDA Model

Yuzhong Chen,Wanhua Li, Wenzhong Guo, Kun Guo*
College of Mathematics and Computer Science
Fujian Key Laboratory of Network Computing and Intelligent Information Processing
Fuzhou University
FuZhou, China
yzchen@fzu.edu.cn, liwanhua_fzu@163.com, fzugwz@163.com, gukn123@163.com

***Abstract*—Micro-blog has become a symbol of the novel social media, and because of its rapid development in such a short time, many research researchers are full of enthusiasm about it. We take use of Latent Dirichlet Allocation (LDA) Model which has excellent dimension reduction capability and can excavate latent semantic from texts to discover popular topics. We improve the original LDA model to FSC-LDA model by combining the text clustering methods and feature selection methods, which can identify the number of topics adaptively. FSC-LDA model can keep short micro-blog texts features better, and make the result more stable. The result of the experiments on real Chinese micro-blog text dataset shows that FSC-LDA model can perform well on the custom evaluation and find more accurate popular topics.**

Keywords—popular topics detection; text clustering; latent dirichlet allocation model; FSC -LDA

I. INTRODUCTION

Micro-blog which attracts many users, gains its explosive development in recent years and has become a novel generation of media nowadays. It is with great significance for monitoring and guiding public opinions to obtain the topics which the Internet users are concerning about from the massive micro-blog texts accurately. So the research on the popular topics detection is necessary.

Compared with Vector Space Model (VSM), LDA model has excellent capacity for finding latent semantic and reducing dimensions. LDA [1] model is a probabilistic model proposed by D.Blei et al in 2003 on the basis of PLSA [2] model, introducing a Dirichlet prior distribution and adding a document layer. This method overcomes the shortcoming of PLSA that the parameters are increased when the number of document set is increasing, and becomes the mainstream of the probability subject model. However, there will be some problems if we achieve the micro-blog topics with LDA model directly. Firstly, when we train the topics by LDA, we have to artificially set the number of topics included in the training documents. In addition, studies in literature [3] show that topic information mined by LDA model will be affected by the length of the input document. Since the micro-blog texts are usually short and the contained information is little, the topic distribution trained by LDA model will deviate from the real topic information. Therefore, if we simply use the LDA model to analyse the micro-blog texts, it will find out many topics with little meanings.

Many researchers improve the LDA model according to different scenarios. Xu et al [4] propose a modified author-topic model to discover users' topics of interest on Twitter. Hong et al [5] put forward an approach to aggregate short texts to long texts according to user relationship or term relationship. Mehrotra R [6] et al aggregate tweets through various pooling schemes before implementing the LDA model to improve topics learned from Twitter content. Although they can solve the problem of the insufficient information of the micro-blog short texts, it cannot conduct incremental learning based on the actual situation, which means when there is a new topic, the topic will not be found with the existing topics-words distributions. Zhang et al dedicate themselves to the Chinese micro-blog topic modeling and propose the MB-LDA model based on LDA [7]. Through the spread relationship of micro-blogs, this model associates short micro-blog texts to form a long text to model topics. Experiments show that this method is superior to the traditional LDA model, but it is more complex and need to be further analyzed on the relationship between micro-blog users. Moreover, it cannot model with the micro-blog texts without being forwarded and commented.

Based on the analysis above, in this paper, we concern about the defects that the traditional LDA model can neither adaptively identify the number of topics nor effectively train the short texts. We proposed FSC-LDA (Feature Selection and Cluster Based on Latent Dirichlet Allocation) model by introducing text clustering method and feature selection method to improve the original LDA model. According to the characteristic of popular topics, we introduce freshness of words and present text feature extraction method called F-TFIDF (Fresh Term Frequency, Inverse Document Frequency). Finally, considering the category distinction degree of words, we propose a category statistic method to extract the keyword sets of topics.

II. TEXT FEATURE EXTRACTIOIN

Due to the short length of the micro-blog texts and less meaningful words contained. Therefore, we need to select keywords from micro-blog texts. Traditional methods are based on word frequency, document frequency of words and the TFIDF value of words to calculate the weight of each word [8]. The method based on word frequency will recognition some pointless and high frequency words as the feature words, compared with the TF method, TFIDF can reduce the weight of

978-1-4673-9372-0/15 $31.00 © 2015 IEEE
DOI 10.1109/WISA.2015.58

this word which with high document frequency. However there will be a problem if we use the above methods directly, the mentioned calculation methods believe that the existing old words are as important as the identified new words, which are adverse to the network new words identification. In this paper, we propose the concept of F-TFIDF value to measure the weight of each word which combines the feature of the micro-blog topics with the traditional TFIDF.

A. Definition of F-TFIDF

Considering the micro-blog topics are usually sudden, and the expression of Chinese micro-Blog is non-mainstream, our paper proposes an improved method for the calculation of short text micro-blog feature extraction which is based on the idea of TFIDF and introduces the freshness of words.

Definition 1 (Word F-TFIDF): F-TFIDF is a new weight mechanism that not only consider TFIDF value of words but also the freshness of the words to extract the text feature, and increase the weight of new words, F-TFIDF is defined as:

$$F\_TFIDF(t,d)=\begin{cases} TFIDF(t,d)+FR(t) & t\in FWS \\ TFIDF(t,d) & t\notin FWS \end{cases} \quad (1)$$

where $FWS=\{NW_1\cdots NW_N\}$ is the set of new words, NW_i is the *i-th* new words, and $FR(t)$ is the freshness of word t. While we are calculating the weight of a word, we have to judge whether the word is contained in the new-words table or not. If the word is in the FWS, we should define the freshness and TFIDF of the word as the weight of this word, or we regard the TFIDF of the word as the weight of this word.

Definition 2(Word Freshness): The frequency of word t is $TF(t)$, and the document frequency of word t is $DF(t)$. The freshness of word $FR(t)$ mentioned above is shown as follow:

$$FR(t)=\frac{TF(t)}{DF(t)} \quad (2)$$

B. Feature Extraction Process

First of all, we use a tokenizer called NLPIR to define all new words from the training micro-blog texts, and add the new words to the new-words table. Secondly, we calculate the F-TFIDF value by the definition above as the weight of words. Finally, we choose the words with higher weight to represent all micro-blog texts.

III. TOPIC DETECTION

LDA model [1] is a symbol of the topic model. For its excellent ability to reduce the dimensions and good scalability, it becomes a popular method in text mining research field recently. It can not only dig latent semantic themes out from massive texts but also be applied to other fields. Against this backdrop, we achieve the popular topics based on micro-blog texts, and make contrastive analysis between the original LDA model and the modified model in accordance with related evaluation.

A. Introduction of LDA Model

LDA model has performed well on the dimension reduction. Besides, it can deal with the problem of the sparse matrix by turning the high-dimensional and sparse documents-words matrix to the smaller-dimensional document-topics matrix. LDA model is based on the bag-words model. It ignores the order of words, which digs out the implicit information and semantics by comparing the common features of the probability distributions on words.

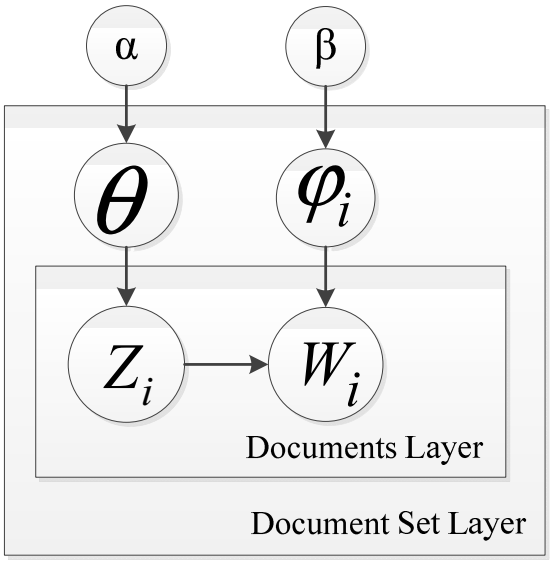

Fig. 1. LDA Model Principle

Fig.1 gives the generation process of each document. α、β are the hyper-parameter and trained from a corpus. α reflects the relative strength of potential theme and β denotes the probability distribution of potential theme. For each document, there will be a θ denoting the probability distribution of each potential topic. φ_i .represents the *i-th* topic' distribution on the words table. Z_i is selected from the θ, and W_i will be defined by Z_i and φ_i .Document d_i is represented as $d_i=\{W_1\cdots W_N\}$ where W_j is the *j-th* word, and N is the scale of the keywords. $D=\{d_1\cdots d_M\}$ is the set of documents , where M is the number of documents. The set of implied themes is $Z=\{Z_1\cdots Z_K\}$, where K is the number of implied-theme.

From Fig.1 we can deduce that generated probabilities of the *j-th* word W_j in the document i can be presented as $P(W_j|d_i)=\sum_{k=0}^{K}P(W_j|Z_k)\cdot P(Z_k|d_i)$, where K is the number of topics, and $P(W_j|Z_k)$ is the probability of W_j in topic Z_K ,

$P(Z_k|d_i)$ is the probability of document di to generate the topic Z_K.

B. FSC-LDA Model

As we known, it is necessary for the original LDA model to assign the number of potential topics before the training. However it is difficult for us to determine the number of topics in actual situations, which requires us to find a useful approach to identify the number of topics automatically.

We propose the LDA-C (Latent Dirichlet Allocation and Cluster) model to do with the problem, which combines the basic text-clustering algorithm with the original LDA model. Besides, the short length of micro-blog texts leads to the frequency of occurrence isn't much enough, and make it difficult to determine whether it is a good correlation between these words or not, which led to the distribution in each topic of each text will be distorted.

We consider that two texts belong to the same topic will be similar in the distributed of potential theme. As a result, we can calculate the similarity of the distributed in potential themes between two texts to cluster the similar text together with some unsupervised clustering methods. Moreover, it is advisable for us to find a number which is much larger than the real topic number. In a certain time frame, the number of popular topics in micro-blog platform is countable and limited. For the distorted distribution of the themes in original LDA model, some measures should be taken on the document-theme matrix to reduce the distortion degree and include more information on the original text before text clustering. Based on the above analysis, this paper presents a FSC-LDA model.

FSC-LDA model draws on the method of feature selection, selecting the feature words from the micro-blog texts to expand the information of micro-blog documents-topics. We take use of the original text feature extraction method to extract the higher-weights words sets of micro-blog texts and build an N-dimensional vector for each micro-blog text, then connect it to the documents-themes matrix which is trained by LDA model. Finally, we can use the text clustering method to find the topics from the treated documents-themes vector. The process of FSC-LDA model is shown in Fig.2.

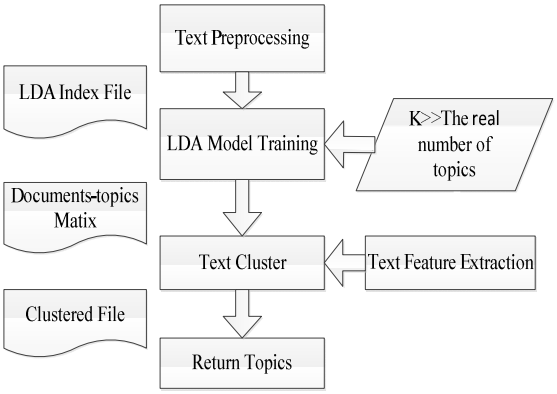

Fig. 2. FSC-LDA Model Framwork

Preprocessing the texts is the first step, when the FSC-LDA model is used to detect the popular topics. In addition, in order to identify the number of topics adaptively, the model choose a number K which is much larger than the real number of topics Kr as the number of topics obtained by LDA model. Then model topics over the input texts by the original LDA model to obtain the documents-topics matrix. Meanwhile, using the text feature extraction methods mentioned above to build the documents-features words matrix. Finally, according to the distribution in the topics and the similarity matrix of feature words, digging out a specific topic clusters by text clustering algorithm.

C. Design and Implementation of Algorithm

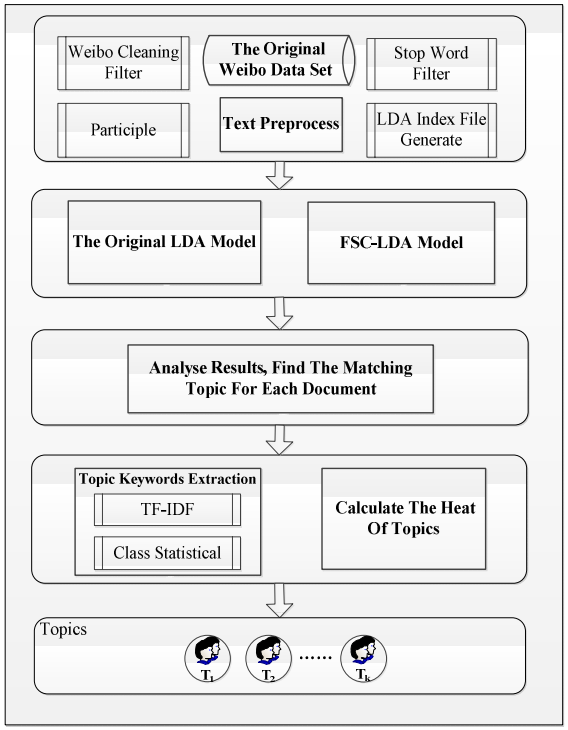

Fig. 3. Popular Topics Detection Framework

Algorithm 1 FSC-LDA Main Algorithm

Input: LDA Index File: ldafile;
Input: The Original Weibo data set: file;
Input: The Topic Number of LDA: K;
Output: Topic Set : T
1. $M1 \leftarrow LDA(ldafile, K)$
2. $M2 \leftarrow featureSelect(file)$
3. $M3 \leftarrow mergeMatix(M1, M2)$
4. $T \leftarrow clusterProcess(M3)$
5. **Return** T

The whole framework of our approach is described as the Fig.3. First of all, we should preprocess the original texts. The content of preprocessing include: micro-blog filter, Chinese word segmentation, stop words filtering, and generation of LDA index files. Then, put the preprocessed data into three

different models to find topics for each micro-blog texts. Finally, we calculate the popularity of each topic to return the popular topics, and extract keywords for the found topics.

Algorithm 2 FeatureSelect(file)

Input: The Original Weibo data set: file
Output: documents-words Matrix : M
4. *weight* ← *0.0*
5. *newWordTable* ← *NLPIR*(file)
6. *TFmap* < *String*, int >← *TFcount*(file)
7. *DFmap* < *String*, int >← *DFcount*(file)
8. *TFIDFmap* < *String*, double >← *TFIDFcount*(*TFmap*, *DFmap*)
9. **For** *w*∈ *file*
10. {
11. **If** (*w*∈ *newWordTable*){
12. *weight* ← *TFIDFmap.get*(w) + *Freash*(w)
13. }
14. **Else**{
15. *weight* ← *TFIDFmap.get*(w)
16. }
17. *weightMap*.put(w, *weight*)
18. }
19. *M* ← *matixGenerate*(weightMap)
20. **Return** *M*

IV. TOPIC EXTRACTION

A. The Definition of Topic Popularity

Definition 3 (Topic Popularity): In order to quantify the activity and influence of the topics, we calculate popularity of topics. Considering that the more micro-blog texts are talked about the topic in a certain time, the greater influence of the topic has. We use the proportion of the total number of micro-blog texts to define the popularity of a topic. So the popularity of topic is defined as:

$$HOT(T_i) = NUM_i \Big/ \sum_{j=0}^{k} NUM_j \tag{3}$$

where $HOT(T_i)$ is stand for the popularity of the *i-th* topic, and NUM_i is the number of micro-blog texts about the *i-th* topic. The popularity of topic can be accurately portrayed by calculating the micro-blog proportion of the total number.

B. Topic Keywords Extraction

In order to reveal the found topics more friendly and intuitively, we need to refine keywords for each topic. With the help of external knowledge HowNet, Tsengt et al. [9] firstly improve the quality of subject description through extracting the relevant keywords, but this method will not be able to respond to new words, and not be suitable for the micro-blog topics extraction. Document [10] proposes an automatic method to identify the keywords of clustered texts based on the LDA model which is according to the documents-themes matrix generated by LDA. An approach named category statistic method is proposed to dig out more reasonable keywords set of topics, which combines traditional statistic methods with the category distinction degree of words.

Definition 4 (Word Distinction Degree): As we know the description on a theme is to help people to distinguish a theme from other themes. As a result, the key phrases to describe topics must have a good category distinction degree. The category distinction degree of word W_i in category C_j is defined as:

$$CD(W_i, C_j) = \frac{TF(W_i, C_j)}{TF(W_i, C_{total})} \quad (W_i \in C_j \cap W_i \in WS) \tag{4}$$

where $TF(W_i, C_j)$ is the TF value of word W_i in category C_j, $TF(W_i, C_{total})$ is the TF value of word W_i in the whole text. WS is the set of alternative words selected by the TFIDF value. The category distinction degree of word W_i in category C_j is measure by the ratio of the frequency of W_i appeared in the category C_j in the frequency of W_i appeared in the whole data. The specific process of category statistic method is shown as follow.

Algorithm 3 Category Statistic Method

Input: Micro-blog on a topic clusters：:categorytexts;
Input: word frequency in total data set: TTFmap
Output: keywords of topic： *keyWords*
1. *TFmap* < *String*, int >← *TFcount*(*categoryposts*)
2. *DFmap* < *String*, int >← *DFcount*(*categoryposts*)
3. *TFIDFmap* < *String*, double >← *TFIDFcount*(*TFmap*, *DFmap*)
4. *WS* ← *TopK*(TFIDFmap)
5. **For** *w*∈ *WS* {
6. *WordDiscrimination* ← *TFmap.get*(w)/*TTFmap.get*(w)
7. *WSmap*.put(w, wordDiscrimination)
8. }
9. *keyWords* ← *TopK*(WSmap)
10. **Return** *keyWords*

Firstly, we calculate the TF value of each word in the whole micro-blog set. Secondly, we calculate the category distinction degree of each alternative word. Finally, pick out the words with high category distinction degree as a set of keywords in the text cluster set.

V. EXPERIMENT

A. Data Sets and Experimental Environment

TABLE I. THE RESULTS OF TWO MODELS ON THE DATA 1

Data Set	source	Texts number	Topics number
Data1	Datatang	4257	16
Data2	Crawled	916,146	9

This paper does experiment on the two data sets, which is described in Table I. Data1 is published in Datatang, which contains 16 manually-annotated topics. Data2 is crawled from lab collection platform from 25 November 2013 to 27 November 2013, and the topics are manually-annotated. In this paper, we use JGibbLDA as LDA modeling platform. It is implemented by Java to support the Windows operating system, which uses Gibbs sampling technique to estimate parameters.

B. Result Evaluation

Evaluation referred to this paper includes precision rate, recall rate and F value which are the integrated indicators of recall and precision rates [11].

Definition 5 (Precision): Precision is an important indicator of the algorithm, which refers to the highest proportion of the hand-classified topics in discovered topics. The precision for the *i-th* topic is defined as:

$$P_i = \max\left(\frac{T_i(M_1)}{T_i}, \quad \cdots, \quad \frac{T_i(M_k)}{T_i}, \quad \cdots, \quad \frac{T_i(M_n)}{T_i}\right) \qquad (5)$$

where T_i is the number of micro-blog texts in the *i-th* topic, and $M_1 \sim M_n$ is the number of hand-classified topics. To the whole model, we define the arithmetic average of total precisions.

Definition 6 (Recall): Recall rate refers to the highest proportion of the discovered topics in hand-classified topics. Suppose that there are n hand-classified topics, and the recall of the *j-th* hand-classified topic is defined as:

$$R_j = \max\left(\frac{M_j(T_1)}{M_j}, \quad \cdots, \quad \frac{M_j(T_k)}{M_j}, \quad \cdots, \quad \frac{M_j(T_m)}{M_j}\right) \qquad (6)$$

where M_j is the number of micro-blog texts in the *j-th* hand-classified topic, $T_1 \sim T_m$ is the discovered topics, $M_j(T_k)$ is the number of the *j-th* hand-classified topic in the *k-th* discovered topics. The overall recall rate is consistent with the precision rate, we marked the arithmetic mean of each topic to the whole algorithm.

Definition 7 (F value): F value is the integration of recall and precision rates, it is their harmonic mean and it is defined as:

$$F = \frac{\left(\beta^2 + 1\right) \cdot P \cdot R}{\beta \cdot P + R} \qquad (7)$$

where β is reconcile parameters which can adjust the precious and recall rates. Experience has shown that the precision is as important as the recall, so $\beta = 1$.

C. Experimental Results Analysis

1) Comparison on Experimental Results

In this section, according to the custom evaluation and comparing the results of manual annotation, we calculate the performance of two models. The results are shown in Table II and Table III.

TABLE II. THE RESULTS OF TWO MODELS ON THE DATA 1

Table Head	Table Column Head Data 1	
	LDA	***FSC-LDA***
Precision	0.7263	0.9260
Recall	0.8518	0.8891
F value	0.7841	0.9071

TABLE III. THE RESULTS OF TWO MODELS ON THE DATA 2

Table Head	Table Column Head Data 1	
	LDA	***FSC-LDA***
Precision	0.8322	0.8729
Recall	0.8605	0.9051
F value	0.8461	0.8887

Table II presents the performance on the first data set of two models. Table III presents the performance on the second data set of two models. We can see that FSC-LDA is superior to the original LDA model on the precision, recall and F values. It can be seen from the comparison of experimental results that the FSC-LDA model proposed in this paper is better the traditional LDA models in discovering the popular topics. Because FSC-LDA model considers the probability distribution and the feature words distribution of text in the topics, which contains relatively more complete information and exploits the information text fully. Not only can it dig out the latent semantic of the text, but also maintain the original features of the text, which solves the problem of poor training results. Therefore, the FSC-LDA model is proposed to discovery the popular topics in short micro-blog texts.

2) Topics Analysis

Our FSC-LDA model uses the K-Means clustering algorithm to cluster the treated matrix and discovers topics. The number of K-Means clustering is determined by DBSCAN clustering algorithm.

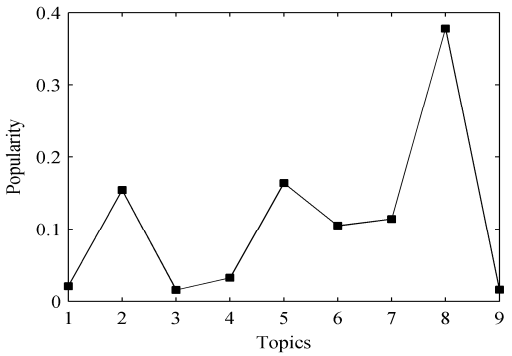

Fig. 4. The popularity distribution of topics on dataset 2

The popularity of the found topics will be analyzed, different keyword extraction methods will be compared to illustrate the advantages of keyword extraction method based on category statistic.

Table IV displays the results of two different keyword extraction methods on the data1. In order to make the result easy to catch for more people, we translate the Chinese contents to English and the corresponding Chinese is shown below to the English one. From the table, it can be found that both methods can perform well in extracting topic keywords. But the proposed method based on the categories of statistic will have better results than the method based on TFIDF directly. The reason is that it can filter out the bad words which have a high value of TFIDF but have little occurrences. Such as the keyword "blingblingsisternamebag" is belong to the keyword set which is extracted in Topic 2 by using the method based on TFIDF, but not belong to the other sets. However, "blingblingsisternamebag" is a word with less meaning. In a word, the method based on category statistic can extract more effective words than the method based on TFIDF.

TABLE IV. THE KEYWORDS OF SOME TOPICS

Topic Number	Artificial label description	Method1 based on TFIDF	Method1 based on category statistic
Topic 0	Land Appreciation Tax 土地增值税	"Real estate Business", "FuLi" , "Many real estate", "find" 房地产商 富力 地产众多 发现	"Real estate Business", "FuLi" , "Many real estate", "huge profits" 房地产商 富力 地产 众多暴利
Topic 1	Two children Topic 二胎话题	"alone", "policy", "open", "population", "children" 单独 政策 放开 人口 孩子	"alone", "policy", "open", "population", "children" 单独 政策 放开 人口 孩子
Topic2	Private custom movies 私人定制电影	"Blingblingsisternamebag", "private", "geocenter", "China", "movie" blingblingsisternamebag 私人 地心 中国 电影	"private", "geocenter", "gravity", "China", "movie" 私人 地心 中国 电影 引力

VI. CONCLUSION

FSC-LDA, the proposed model considering the freshness of the words in the micro-blog texts feature extraction to expand the obtained documents-themes matrix and constructs the mixing matrix. Then we cluster the mixing matrix with the text clustering algorithm to discovery topics. Compared to the traditional LDA model, FSC-LDA model can not only retain the advantage of LDA model but also identify the number of topics adaptively in the documentation set, which is suitable for short text topics discovery. In the topic keyword extraction, this paper introduces the category distinction degree of words. The proposed method based on the category statistic can extract a more reasonable set of keywords than other methods.

The clustering algorithm involved in the proposed model is traditional. In other words, it has not yet taken the partial update of data sets in practical application into account, which may bring large overhead re-clustering. In addition, only the number of topics is considered in the definition of topics popularity, other influential factors in micro-blog are not considered. We will try to use incremental clustering algorithm to find the topics, and do further analysis on the found topics.

ACKNOWLEDGMENTS

The authors would like to thank the support of the Program of National Natural Science Foundation of China under Grant No.61300104 and No.61103175, the Key Project of Fujian Education Committee under Grant No.JK2012003, the Technology Innovation Platform Project of Fujian Province under Grant No.2009J1007, the Key project of Industry-Academic Collaboration of Fujian Province under Grant No.2014H6014 and the Natural Science Foundation of Fujian Province under Grant No.2013J01230 and No.2013J01232.

REFERENCES

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 (2003): 993-1022.

[2] Hofmann, Thomas. "Unsupervised learning by probabilistic latent semantic analysis." Machine learning 42.1-2 (2001): 177-196.

[3] Lu, Yue, and Chengxiang Zhai. "Opinion integration through semi-supervised topic modeling." Proceedings of the 17th international conference on World Wide Web. ACM, 2008.

[4] Xu Z, Lu R, Xiang L, et al. Discovering user interest on twitter with a modified author-topic model[C]//Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on. IEEE, 2011, 1: 422-429.

[5] Hong, Liangjie, and Brian D. Davison. "Empirical study of topic modeling in twitter." Proceedings of the First Workshop on Social Media Analytics. ACM, 2010.

[6] Mehrotra R, Sanner S, Buntine W, et al. Improving lda topic models for microblogs via tweet pooling and automatic labeling[C]//Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013: 889-892.

[7] Zhang, Chenyi, Jianling Sun, and Yiqun Ding. "Topic mining for microblog based on mb-lda model." Journal of Computer Research and Development48.10 (2011): 1795-1802.

[8] Yatsko V, Dixit S, Agrawal A J, et al. TF* IDF Revisited[J]. intelligence, 2013, 16(4): 2.

[9] Tseng, Yuen-Hsien, et al. "Toward generic title generation for clustered documents." Information Retrieval Technology. Springer Berlin Heidelberg, 2006. 145-157.

[10] Jing Shi, Wanlong Li. "Topic Words Extraction Method Based on LDA Model." Computer Engineering, 2010, 36(19).

[11] Yiling Zeng, Hongbo Xu, Shuo Bai. "Research on the extraction and organization of key phrases in Web texts." Journal of Chinese Information Processing, 2008, 22(3): 64-70, 80.