ICIC Express Letters
Part B: Applications
Volume **7**, Number **3**, March **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **621–625**

HOT TOPIC CLASSIFICATION OF MICROBLOGGING BASED ON CASCADED LATENT DIRICHLET ALLOCATION

Jianfeng Fu, Nianzu Liu, Cuihua Hu and Xujie Zhang

College of Mathematics and Information
Shanghai Lixin University of Commerce
No. 2800, Wenxiang Road, Shanghai 201620, P. R. China
{ fujianfeng; liunianzu; hucuihua; zhangxujie }@lixin.edu.cn

Received September 2015; accepted December 2015

Abstract. *Short text of microblog is a serious challenge to traditional text classification methods. To address this problem, this paper presents a novel model of cascaded (two layers) latent Dirichlet allocation for hot topic classification of microblogging. The model integrates the reply and retweet messages as well as the microblogging content for representing the feature space of the microblog. The first latent Dirichlet allocation is used to find the messages posted by reply and retweet which are closely related to microblogging original content, and the second latent Dirichlet allocation is employed to model microblog with the content and the content-related messages. Experimental results demonstrate the outstanding performance of our model by comparing with TF-IDF and traditional latent Dirichlet allocation model.*
Keywords: Hot topics, Microblogging, Cascaded latent Dirichlet allocation

1. **Introduction.** With the development of Web2.0 and online social networks, microblog has attracted widespread attention. According to Twitter, there were 645 million active registered Twitter users by the end of July, 2014. Microblog is a platform based on social network where users can send and receive messages. However, for the limitation of microblog, each message is not allowed to be longer than 140 characters. This is really a big challenge for Natural Language Processing (NLP) to parse, classify and even understand the short messages.

Many different research efforts for instance, feature enrichment and selection, have been exploited for mining the microblogs [1,2]. The previous work mainly focuses on using external resource (for exmaple, WordNet and Wikipedia) for expanding the short text of microblogs. However, the systems performance may deeply depend on the quality of the external resource and it is hard to find a universal applicable external resource. To address this problem, some other researches focus on using latent Dirichlet allocation (or LDA for short) model [3] and its extension to extract the hidden topics information for mining the microblogs [4]. Ramage et al. [5] employed label LDA, which extends LDA and maps the content of Twitter feed into four dimensions of substance, social, status and style, to classify the content of the Twitter and the users. Zhao et al. [6] developed a Twitter-LDA model to detect topics from Twitter. Different from traditional topic models, the Twitter-LDA takes a single tweet as a single topic. The experimental results have shown the outstanding performance of Twitter-LDA. Zhang and Sun [7] introduced an improved LDA model of MB-LDA, which takes both contactor and document relevance relation into consideration to improve topic mining in microblogs. Not only the hidden topics of microblogs but also the topics focused by contactors can be discovered with the model. Mehrotra et al. [8] examined twitter-pooling schemes to improve LDA topic quality and proposed a novel scheme of hash-tag based pooling. The pioneer studies show

that the LDA and its extensions are powerful tools and provide a more convenient and efficient way for resolving the problem of short text of microblog.

Microblog allows users to exchange content by reply and retweet function. An interesting topic on the microblog will become hot as quickly as possible while thousands of people retweet and reply it. Most of the users are in discussions around the topic, and their messages are closely related to the content of the topic. The content-related messages are supplement or extension of the topic and will be helpful to overcome the shortage of short text of microblog. In this paper, we focus on modeling and classifying microblogging hot topics. The main contributions of this paper are as follows.

(1) We propose a strategy which integrates the content-related messages of reply and retweet as well as the microblogging content to represent the feature space of the microblog.
(2) We present a model of cascaded LDA for modeling microblogging hot topics. It mainly involves two stages. In the first stage, we use LDA to find the messages of reply and retweet which are closely related to the microblogging content. In the second stage, we employ LDA to model the microblog with the content and content-related messages.

The remainder of the paper is organized as follows. In Section 2, we introduce our system framework. In Section 3, we give a detailed description of the presented cascaded LDA. In Section 4, we report the dataset, empirical experiments and comparison results. In Section 5, we conclude and present the future work.

2. **System Framework.** We now introduce the overview of the whole processing that aims to classify microblogging hot topics. Above all, our objective is Sina Weibo, the most popular microblogging service in China. In the step of pre-processing, we segment the Chinese words and remove stop words. We also design a spam filter to remove the microblogging spam. Microblogging spam is unsolicited, repeated or unwanted messages or links which can impact other users negatively. In this paper, we focus on the microblogging spam posted by reply and retweet function of hot topics. Both "keywords" based and similarity based spam filters are involved in the strategy. Since the microblogging spam is with obvious literally characteristic, we design a "keywords" based rules set to detect them. Besides, we use "messages similarity" based rule to detect similar messages. The greater the similarity of two messages is, the larger the possibility that they are repeated messages is. After that, we employ the first LDA model to find the top N messages of reply and retweet which are closely related to microblogging original content. In the second LDA model, we integrate the microblogging content and content-related messages for representation of the semantic features of microblog. Finally, we use classifiers to evaluate the presented method. The system framework is shown in Figure 1.

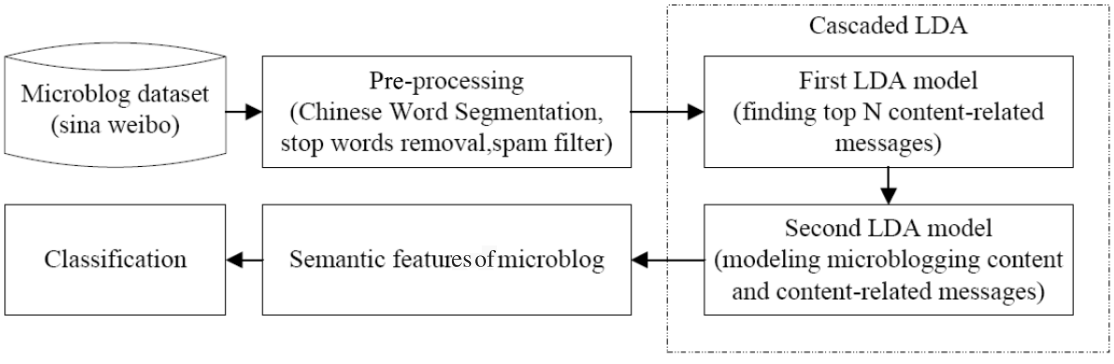

FIGURE 1. The system framework

3. **The Cascaded LDA.** Latent Dirichlet allocation (LDA) is a generative probabilistic model for collections of discrete data [3]. It is a topic model and often used to model relations between topics. In this paper, we propose a cascaded LDA method to model the microblogging hot topics.

3.1. **The first LDA.** In the first LDA of this study, we consider each message posted by reply and retweet function in a microblog as a single document. By using "bag of words" approach, each message (including the microblogging original content) is treated as a vector of word counts. It is represented as a probability distribution over some latent topic, while each topic is represented as a probability distribution over some words. Formally, we assume that there are T latent topics and M messages in a hot microblog, and L is length of each message. Let θ_m denote the topic distribution for message m, and ϕ_t denote the word distribution for topic t. Let $Dir(\alpha)$ denote the Dirichlet distribution for parameter α. The generation progress of first LDA is described in Figure 2.

For each microblog do
  For each message $m = 1,...,M$ do
    Draw $\theta_m \sim Dir(\alpha)$
  End for
  For each topic $t = 1,...,T$ do
    Draw $\phi_t \sim Dir(\beta)$
  End for
  For each of the word position m, l do
    Choose a topic $t_{m,l} \sim Multinomial(\theta_m)$
    Choose a word $w_{m,l} \sim Multinomial(\phi_{t_{m,l}})$
  End for
End for

FIGURE 2. The generation progress of the first LDA

Given the parameters α and β, the joint distribution of the model is given by:

$$P(\theta, T, M|\alpha, \beta) = P(\theta|\alpha) \prod_{i=1}^{L} P(t|\theta) P(w_i|t, \beta) \tag{1}$$

Learning the various distributions is a problem of Bayesian inference. In this paper, we use Gibbs sampling to infer the model. As a result of Gibbs sampling for the first LDA on each microblog, θ is an $M * T$ matrix. In the matrix, each line is a message of a microblog and each element value indicates the implicit topic of a word. Let M_0 denote the original content of a microblog, and M_i, $i = 1, \ldots, n$ denote the reply and retweet messages. It is easy to calculate $\cos(M_0, M_i)$ (the Cosine similarity of M_0 and M_i) and find top N messages related to the original content of the microblog.

3.2. **The second LDA.** By accomplishment of the first LDA, we obtain top N related messages posted by reply and retweet function in each microblog. Different from the first LDA, we consider the integrated content (the microblogging original content and the top N content-related messages) as a single document in this stage. The document is finished by not only the original author but also the repliers and retweeters. They are a group of authors. When writing a microblog, the authors first choose a topic based on topic distribution. Then they choose a bag of words one by one based on the word distribution of chosen topic. The generation progress of the second LDA is similar to the first LDA. Due to the space limitation, we leave it out.

4. Experiments and Results.

4.1. Dataset.

We designed a web crawler to collect the hot topics from Sina Weibo for evaluation. The collected hot microblogs were published from May 1, 2014 to August 1, 2014. A microblog is considered as not hot and excluded in the collection if the total number of replies and retweets is less than 100. We removed microblogs which had fewer than 5 words. Each microblog in the collection was manually assigned into one of seven predefined categories: Entertainment, Business, Sports, Health, Education, Travel and others. We use the first six categories which contain 155 hot microblogs and 922721 replies and retweets (in order to improve system performance, we use the first 1000 replies and retweets of a microblog if the messages number is over 1000) for our experiment. The detail information of the collection is shown in Table 1. We use 5-fold cross validation to evaluate the classification performance.

Table 1. Detail information of dataset

	Entertainment	Business	Sports	Health	Education	Travel
microblogs	326	318	305	220	206	175
replies & retweets	226359	190265	177832	118172	125743	84350

4.2. Performance comparison.

We conduct the experiments to evaluate the proposed model of cascaded LDA by comparing with TF-IDF and traditional LDA model. Two classifiers, Support Vector Machine (SVM) and Naïve Bayesian (NB), are employed to demonstrate the universality of the model. We use multinomial Bayesian classifier[1] and LIBSVM [9] with linear kernel function in our experiments. Several widely-used performance metrics are utilized to evaluate the classification task: precision (P), recall (R), F1 score and micro-average. Table 2 and Table 3 display the comparison results of TF-IDF, LDA and cascaded LDA by using SVM and NB. Here, based on preliminary experiments, we set the values of the parameters as $\alpha = 1$, $\beta = 0.1$, $T = 50$ for traditional LDA and the cascaded LDA, and $N = 30$ for finding the top N content-related messages.

Table 2. Classification performance comparison based on SVM

	TF-IDF			LDA			Cascaded LDA		
Category	P	R	F1	P	R	F1	P	R	F1
Entertainment	0.6779	0.6653	0.6715	0.6580	0.7014	0.6790	0.8257	0.8010	0.8132
Business	0.6496	0.6786	0.6638	0.6951	0.6482	0.6708	0.8277	0.7888	0.8078
Sports	0.6101	0.6291	0.6195	0.6401	0.6761	0.6576	0.8389	0.8116	0.8250
Health	0.6377	0.6015	0.6191	0.6379	0.6480	0.6429	0.7618	0.7976	0.7793
Education	0.6135	0.6692	0.6401	0.6253	0.6414	0.6332	0.7816	0.8124	0.7967
Travel	0.6678	0.6791	0.6734	0.6561	0.6320	0.6438	0.7930	0.8274	0.8098
Micro-average	0.6433	0.6539	0.6483	0.6547	0.6621	0.6579	0.8101	0.8046	**0.8070**

Table 2 presents the comparison results based on SVM. It can be seen that our proposed model achieves the best classification performance on each category and makes a significant improvement on micro-average F1 score as well as the comparison results based on NB which is shown in Table 3. The comprehensive comparison results demonstrate that our proposed model performs better than TF-IDF and LDA on microblogging hot topics classification. It is due to the fact that the integrated messages of reply and retweet enrich the text representation and enhance the classification performance of microblogs. In addition, it can be observed that the LDA model does not achieve outstanding performance (less than 0.01 on micro-average) while comparing with TF-IDF model on microblogging

[1]http://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html

TABLE 3. Classification performance comparison based on NB

	TF-IDF			LDA			Cascaded LDA		
Category	P	R	F1	P	R	F1	P	R	F1
Entertainment	0.6853	0.6708	0.6780	0.6123	0.6802	0.6445	0.7699	0.8262	0.7971
Business	0.6477	0.6453	0.6465	0.6959	0.6723	0.6839	0.8629	0.8021	0.8314
Sports	0.6348	0.6224	0.6285	0.6796	0.6214	0.6492	0.7781	0.7922	0.7851
Health	0.6236	0.6145	0.6190	0.6159	0.6054	0.6106	0.8068	0.7683	0.7871
Education	0.6178	0.6359	0.6267	0.7233	0.6567	0.6884	0.7891	0.7608	0.7747
Travel	0.6548	0.6771	0.6658	0.6528	0.6461	0.6494	0.8075	0.8144	0.8109
Micro-average	0.6465	0.6441	0.6452	0.6625	0.6494	0.6551	0.8026	0.7963	**0.7989**

hot topics classification. A reasonable explanation of this result is that the short text of microblog leads to low performance of traditional LDA to model latent topics.

5. **Conclusions.** In this paper, we proposed a novel model of cascaded LDA to classify hot topics of microblogging. The model solves the text sparseness problem of microblog by integrating the reply and retweet messages as well as the microblogging original content for representation of the feature space of the microblogs. Experimental results demonstrated the outstanding performance of our model by comparing with TF-IDF and LDA. Future work involves: (1) using more advanced techniques to filter microblogging spam efficiently and (2) improving the strategy of evaluating content-related messages of reply and retweet.

Acknowledgment. This work is supported by Innovation Program of Shanghai Municipal Education Commission (14YZ151) and China National Social Science Foundation (13CTQ042). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Z. Liu, W. Yu, W. Chen, S. Wang and F. Wu, Short text feature selection for micro-blog mining, *International Conference on Computational Intelligence and Software Engineering (CiSE)*, pp.1-4, 2010.

[2] J. Tang, X. Wang, H. Gao, X. Hu and H. Liu, Enriching short text representation in microblog for clustering, *Frontiers of Computer Science*, vol.6, pp.88-101, 2012.

[3] D. M. Blei, A. Y. Ng and M. I. Jordan, Latent Dirichlet allocation, *The Journal of Machine Learning Research*, vol.3, pp.993-1022, 2003.

[4] L. Hong and B. D. Davison, Empirical study of topic modeling in Twitter, *Proc. of the 1st Workshop on Social Media Analytics*, pp.80-88, 2010.

[5] D. Ramage, S. Dumais and D. Liebling, Characterizing microblogs with topic models, *ICWSM*, vol.10, pp.130-137, 2010.

[6] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan et al., Comparing Twitter and traditional media using topic models, *Advances in Information Retrieval*, pp.338-349, 2011.

[7] C. Zhang and J. Sun, Large scale microblog mining using distributed MB-LDA, *Proc. of the 21st International Conference Companion on World Wide Web*, pp.1035-1042, 2012.

[8] R. Mehrotra, S. Sanner, W. Buntine and L. Xie, Improving LDA topic models for microblogs via tweet pooling and automatic labeling, *Proc. of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.889-892, 2013.

[9] C.-C. Chang and C.-J. Lin, LIBSVM: A library for support vector machines, *ACM Trans. Intelligent Systems and Technology (TIST)*, vol.2, p.27, 2011.