# *Research on Campus Hot Topic Detection Based on LDA Topic Model*

**Xiujuan Yi [1, a], Weidong Zhu[2, b]**

[1]*School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044, China*
[2]*Information Center, Beijing Jiaotong University, Beijing, 100044, China*
[a]*16120443@bjtu.edu.cn*, [b]*wdzhu@bjtu.edu.cn*

***Keywords:*** Text modeling; LDA topic model; VSM model; TF-IDF; JS; k-means;

***Abstract:*** The hot topics in colleges and universities network are detected by analyzing students' Internet content. The students' Internet contents have the characteristics of varying lengths, scattered topics and distorted information. The traditional VSM model calculates the weight forming vector of features according to the word frequency statistics and ignores the implicit content in the text. Therefore, the LDA (Latent Dirichlet Allocation) topic model that can identify the hidden topics in the document is used to reconstruct the model and detect topics. When preprocessing text, the text is filtered based on TF-IDF after word segmentation, and then the text is modeled by LDA. After LDA clustering, the k-means algorithm based on the JS (Jensen-Shannon) distance function is used to cluster the documents according to the probability distribution of the subject two times to get the students' Internet theme distribution.

## 1. Introduction

According to the 41st Statistical Report on China's Internet Development Status released by China Internet Network Information Center (CNNIC) in 2018[1], as of December 2017, the number of Internet users in China reached 772 million, and the penetration rate reached 55.8%.

In recent years, with the expansion of the network scale of colleges and universities, the campus network plays an important role in school education as an important field in the Internet. In the structure of netizens in China, the main users of the campus network are students. The life and learning of students has been fully penetrated by the network, affecting the formation of their views of life, value, and world. Therefore, obtaining students' attention topics helps the school to quickly identify and track hot topics and abnormal topics on campus, and take effective measures to guide them. The discovery of hot topics stems from the modeling of text. The traditional text modeling method is the VSM model [2], which uses word frequency to create text vector. However, the VSM model ignores the potential semantic associations between texts. Therefore, this paper uses the information extracted from the student's online log to crawl the corresponding web content. The

content of the webpage is segmented and labeled, and the TF-IDF weight is used to filter words whose weight is too low. The LDA topic model is used to reconstruct text model and to calculate the subject probability vector for students' internet content. The traditional text clustering algorithm uses cosine similarity calculation. In the LDA theme model, the text is the subject probability vector that obeys the Dirichlet distribution, and the similarity uses the JS distance function to be more advantageous. The experimental results show that the LDA text modeling method improved by preprocessing and similarity calculation proposed in this paper is more accurate than the traditional VSM modeling method.

## 2. Methods

### 2.1 Topic Discovery Research

At present, the research on topic discovery mainly focuses on social network topic discovery and news hot topic discovery. The topic is detected from news pages or blogs, and the research on topic discovery based on campus network is still relatively small. Yang Xiang [3] proposed a text segmentation technique based on trie tree, a feature selection algorithm based on tf-idf and a short text clustering algorithm based on improved K-means algorithm for clustering of short text topics. ZHU Xiaofeng et al [4] improved the traditional clustering algorithm by changing the initial clustering center to perform topic discovery. Lai Jinhui et al [5] improves the traditional algorithm by means of isolated point processing. However, these traditional methods lose a lot of deep semantic information and reduce the accuracy of topic discovery.

Therefore, in order to improve the limitations of the above methods, the researchers proposed a topic model, which is a modeling method for implicit topics in these words. In the topic model, the relationship between documents and topics can be represented by generating model. The topic model is obtained by training the text-word matrix to get the text-topic matrix and the topic-word matrix. The probability of each word in the document can be expressed as:

$$p(word \mid doc)=\sum_{topic} p(word \mid topic) \times p(topic \mid doc) \tag{1}$$

Common topic models are PLSA (Probabilistic Latent Semantic Analysis) and LDA.

### 2.2 LDA Topic Model Research

The LDA topic model [7] is an unsupervised machine learning method that can be used to identify topic information implied in a document set by constructing a feature word, an implicit topic, and a three-layer Bayesian model of the document. In the LDA topic model, the topics in the implicit topic set are represented by a collection of feature words and all documents in a document set can be thought of as being composed of a set of topics in an implicit topic set with a certain probability. The three-layer topology of the LDA topic model is shown in Figure 1.

In the LDA topic model, the implicit topics of a document are randomly distributed with different probabilities [8]. The LDA topic model uses the word bag method. The word bag method refers to representing a document as a vector combination containing all words, so that the text can be transformed into a mathematically expressed vector, which is easy to model. The word bag method does not consider the order between words and words in the document, which simplifies the complexity of the model representation and makes it easy to calculate.

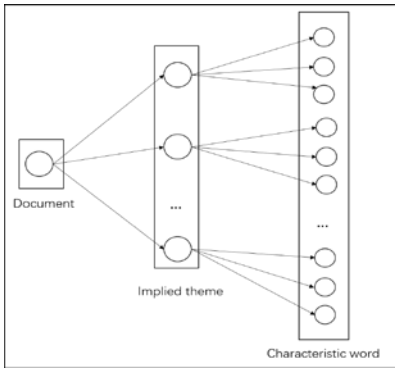


Fig.1 LDA three-layer topology

## 3. Design of Campus Network Hot Topic Discovery

### 3.1 LDA Topic Modeling

The text generation process of the LDA topic model is shown in Figure 2. The LDA model generates a series of observation words $W_{m,n}$ and the words are assigned to the text $\vec{W}_{\mathrm{m}}$. For each document, the mix ratio $\vec{\theta}$ is extracted and the specified words are extracted from the word distribution of the topic. For each word, the topic number $\mathrm{z}_{m,n}$ is sampled based on the mix ratio and the word is extracted by the corresponding topic specified words distribution $\vec{\varphi}_{z_{m,n}}$.

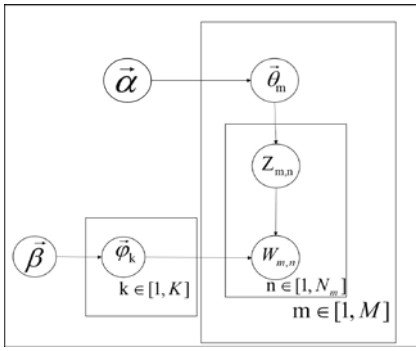


Fig.2 LDA directed probability map

Table 1 LDA Model Parameter Description

| parameter | Description |
|---|---|
| $\vec{\alpha}$ | Dirichlet distribution super-parameter, K-dimensional vector |
| $\vec{\theta}_{\mathrm{m}}$ | The subject mix ratio of the document m, K-dimensional vector; $M \times K$ matrix |
| $\vec{\beta}$ | Dirichlet distribution super-parameters, V-dimensional vectors |
| $\vec{\varphi}_{\mathrm{k}}$ | Mixed component of topic k (word distribution), V-dimensional vector; $K \times V$ matrix |
| M | Number of documents generated, scalar |
| K | Number of topics, scalar |
| V | Number of dictionary words scalar |
| $N_{\mathrm{m}}$ | The length of the document m, extracted by the Poisson distribution with the parameter $\xi$ |
| $Z_{m,n}$ | The mark of the subject selected by the nth word of the document m |
| $W_{m,n}$ | The mark of the nth word of the document m |

### 3.2 Parameter Estimation

The LDA model is mainly based on the following inferences: (1) word distribution $p(w_n \mid z_k = k) = \vec{\varphi}_k$ for each topic $z_k$ ; (2) topic distribution $p(z_k \mid d_m = m) = \vec{\theta}_m$ for each text $d_m$ . The parameters $\Phi=\{\vec{\varphi}_k\}_{k=1}^{K}$ and $\Theta=\{\vec{\theta}_m\}_{m=1}^{M}$ to be solved can be calculated by methods such as Expectation Maximum (EM), Expectation Propagation (EP) and Gibbs Sampling (GS). This paper uses Gibbs sampling estimation. Compared with the other two methods, the Gibbs sampling algorithm has advantages in accuracy and processing speed compared to the other two methods. The Gibbs sampling is one of Markov Chain Monte Carlo (MCMC), whose main idea is to change only one dimension at each iteration until the parameters of the convergent output are estimated. In the LDA model, the dimension is the size of the vocabulary set, and each iteration estimates the topic probability of the current word based on the topic assignment of other words. The full conditional distribution of the topic for a word number i=(m,n):

$$p(z_i = k \mid \vec{z}_{\neg i}, \vec{w}) \propto \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t} \cdot \frac{n_{m,\neg i}^{(t)} + \alpha_k}{[\sum_{k=1}^{K} n_m^{(k)} + \alpha_k]\text{-}1} \tag{2}$$

The steps of the Gibbs sampling algorithm are as follows:

1) Select the appropriate topic number K, hyperparameter vector $\vec{\alpha}$ , $\vec{\beta}$ .
2) Each word of each document in the corresponding corpus is randomly assigned a topic number z.
3) Rescan the corpus, each word using Gibbs samples to update its topic number and the number of the word in the corpus.
4) Repeat the Gibbs sampling based on the coordinate rotation of step 2 until the Gibbs sample converges.
5) Document topic distribution $\vec{\theta}_m$ is derived by counting the subject of each word for all documents. And the distribution of LDA topics and words $\vec{\varphi}_k$ is derived from all the subject terms in the statistical corpus.

### 3.3 Improved cosine similarity calculation

The topic distribution of the students' internet content is obtained by solving the LDA topic model. The cosine similarity clusters the documents as follows:

$$Sim_{LDA}(d_i, d_j) = \frac{\sum_{t=1}^{k} (d_{it} * d_{jt})}{\sqrt{\sum_{t=1}^{k} d_{it}^2} * \sqrt{\sum_{t=1}^{k} d_{jt}^2}} \tag{3}$$

Where $d_{it}$ , $d_{jt}$ represent the probability of the t-th topic of the i, j documents, respectively, and k represents the number of topics. The topic probability vector of the LDA model is subject to the Dirichlet distribution, and the result of the cosine similarity clustering is not accurate. Therefore, the JS similarity distance function capable of measuring the probability distribution distance is used to define the topic probability vector. Expressed as follows:

$$Sim_{LDA}(\vec{d}_i, \vec{d}_j) = 1/2(\sum_{t=1}^{k} d_{it} \ln \frac{d_{it}}{d_{jt}} + \sum_{t=1}^{k} d_{jt} \ln \frac{d_{jt}}{d_{it}}) \tag{4}$$

## 4. Simulation Experiment

### 4.1 Data source and processing

The data source of this paper is the two-day online access record of the students in the city hotspot billing system. After pre-processing of webpage data, a total of 30,000 effective webpages of 300 accounts are randomly selected, and webpage crawling technology is used to crawl webpage content. The webpage is cleaned to remove useless information. Finally, each document is segmented and the stop words are removed. The Jieba is used for text segmentation and part-of-speech tagging. The document-word matrix is constructed by text vectorization to prepare for the next experimental analysis.

### 4.2 Text clustering based on LDA model

First, the text is filtered using the TF-IDF algorithm, and then the Gibbs sampling algorithm is used to solve the LDA model. The $\alpha$ =50/K, $\beta$ =0.01 are assigned in the parameter estimation process. The optimal number of topics K=30 is calculated to obtain a topic-word distribution table. Expressed as follows:

$$\frac{1}{P(w|K)} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{P(w|z^{(m)})} \tag{5}$$

Where M is the number of samples. When the value of $P(w|K)$ is larger, the result of the LDA model clustering the text set is more accurate. Therefore, the optimal number of topics can be judged by the value of $P(w|K)$.

According to the clustering results of LDA, some themes still have great similarities, so the JS distance function is used to calculate the similarity of the documents, and the new document clustering results are obtained. According to the results of LDA clustering, the boundaries of some topics are not clear enough and can be merged. The topic distribution of some documents is shown in Figure 4. Some documents have the same distribution probability on both topics.

Fig. 4 Document topic distribution (partial)

The k-means clustering algorithm is used to cluster the obtained topic probability vectors, and 10 topics are selected from 30 topics for clustering.

### 4.3 Text clustering based on SVM model

The basic idea of vector space model VSM is to transform the processing of text into mathematical vector calculation.The i-th feature item in the web page is represented by $T_i$, and the weight of the feature item is represented by $W_i$, and the web page document can be represented by the vector $\vec{V}(d)=\{w_1,w_2\cdots w_n\}$. The web page feature vector weight $W_i$ is calculated by TF-IDF, and the formula is as follows:

$$\text{TF-IDF=tf}*\text{idf}_i=\frac{n_{ij}}{\sum_k n_{ij}}*\log\frac{|D|}{|\{d:d\in t_i\}|} \tag{6}$$

Where $n_{ij}$ represents the number of occurrences of feature item $t_i$ in web document $d_i$, and the denominator is the sum of the occurrences of all feature items in document $d_j$. $|D|$: is the total number of feature words in all documents, and the denominator is the number of documents containing the feature word. Thus, the similarity to the content of the web page is converted into a vector calculation. In this paper, the k-means clustering algorithm is used to cluster texts. The similarity calculation method is a cosine similarity calculation method.

### 4.4 Evaluation of experimental results

The evaluation methods of clustering results include internal evaluation method, external evaluation method and relative evaluation method. In this paper, the external evaluation method, namely F value, is used to evaluate. The F value is to evaluate the quality of clustering results by using Recall Rate and Precision Rate. The calculation method is as follows:

$$F(\text{i},j)=\frac{2\times P\times R}{P+R} \tag{7}$$

$$P=precision(i,j)=\frac{Nij}{Ni} \tag{8}$$

$$R=\text{recall}(i,j)=\frac{N_{ij}}{N_j} \tag{9}$$

Where $N_i$ indicates that the number of samples whose class is marked i in the original data set, and $N_j$ indicates the number of objects whose class is marked j in the clustering result. $N_{ij}$ is the number of samples contained in the intersection of the cluster with the class i in the data set and the cluster with the class j in the clustering result.

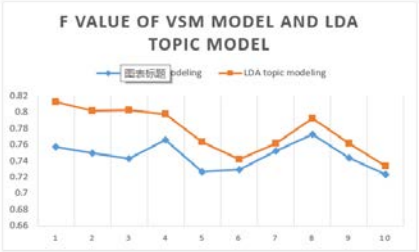


Fig.5 F value line chart

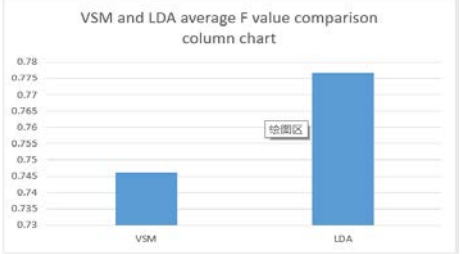


Fig. 6 F-value column chart

In this paper, 200 papers were randomly selected as test sets to evaluate the effect of clustering.

The traditional VSM space vector model is compared with the LDA topic model clustering algorithm based on improved similarity calculation. The F-value comparison between the VSM model based on cosine similarity and the LDA model based on JS similarity is shown in Figure 5. The accuracy results of the subject distribution of the LDA model and the VSM model are shown in Figure 6.

## 5. Conclusion

The experimental results show that the LDA topic model has better clustering effect than the traditional VSM space vector model when the text length is different and the topic is scattered. Moreover, the LDA topic model can discover potential topics and pay close attention to students' internet topics, which helps the school manage students' internet behavior. In the LDA topic model, the JS distance function to calculate text similarity is better than the cosine similarity clustering. However, when short text documents increase, the effect of clustering will become worse. In this respect, the method remains to be improved.

## References


*[1] http://www.cunic.net.cn/hlwfzyj/hlwxzbg/hlwtjbj/201803/t 20180305_70249.html*

*[2] M.Ikonomakis, S.Kotsiantis, V.Tampakas. Text Classification Using Machine Learning Techniques [J]. Wseas Transactions on Computers, 2005, 4(8): 966-974.*

*[3] Yang Xiang. Algorithm and application design and implementation of cluster analysis for short text data [D]. Beijing University of Posts and Telecommunications, 2014.*

*[4] ZHU Xiaofeng, CHEN Chuchu, YIN Yijuan. Research on Improvement of K-Means Algorithm Based on Weibo Public Opinion Monitoring [J]. Information Theory & Practice, 2014, 37(1): 136-140.*

*[5] Lai Jinhui, Liang Song. A Method for Discovering Hot Topics of Microblogs to Eliminate Isolated Points [J]. Journal of Computer Applications and Software, 2014(1):105-107.*

*[6] Zhao Yan, Zhou Bin, Chen Ruhua. Research on Text Classification Algorithm [J]. Software Guide, 2013, 12(10): 54-56.*

*[7] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. J Machine Learning Research Archive, 2003, 3:993-1022*

*[8] Wang Shaopeng, Peng Yan, Wang Jie. Application of LDA-based text clustering in network public opinion analysis[C]// China Conference on Trusted Computing and Information Security. 2014.*

*[9] Wang C, Blei D, Heckerman D. Continuous Time Dynamic Topic Models[C]// Proc. Conference on Uncertainty in Artificial Intelligence. 2012:579--586..*

*[10] Alsumait L, Barbara D, Domeniconi C. On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking[C]// Eighth IEEE International Conference on Data Mining. IEEE Computer Society, 2008:3-12.*

*[11] Fu Ling, Zhang Hui. Multi-document summary combining LDA and spectral clustering [J]. Computer Engineering and Applications, 2013, 49(16): 142-145.*