Intelligent Computing Technology and Automation
Z. Hou (Ed.)

doi:10.3233/ATDE231269

Text Semantic Analysis Algorithm Based on LDA Model and Doc2vec

Weiwei ZHANG[a], Guangyu ZHAI[b.1], Binbin ZHONG[a], Xiaoyi KONG[a]

[a] *Gansu Provincial Meteorological Information and Technical Equipment Support Center, Lanzhou, Gansu 730020, China*
[b] *School of Economics Management, Lanzhou University of Technology, Lanzhou, Gansu 730050, China*

Abstract. With the rapid development of the Internet, massive messy text data is distributed in all walks of life, how to quickly and effectively mine meaningful semantic information from these complex and messy texts has become an important task in the field of natural language. First, texts such as Weibo comments lack contextual semantics due to sparse data, which in turn affects the effect of text semantic mining. Second, existing topic model use semantic representations in different vector spaces, resulting in low accuracy. Therefore, this paper proposes a new text semantic analysis model: DBOW-LDA model, which integrates the LDA topic model (Latent Dirichlet Allocation) and the sentence vector model (Doc2vec), so that the output contains the given topic semantic information. Sentence vector representation of text. The accuracy of the algorithm is further improved. In the experiment, the crawled Weibo comment text is used as the data set, and the K-Means clustering algorithm is used to compare the effects of each model. The experimental results show that the clustering effect based on DBOW-LDA model is better than Word2vec, WT-LDA, LDA+Word2vec model.

Keywords. Semantic Analysis; Topic Model; LDA; Doc2vec

1. Introduction

The rapid development of the Internet has resulted in the rapid dissemination of information on social media. People have formed public opinions based on some opinions, views and opinions expressed on hot topics. Textual semantic mining of a large amount of text information can obtain public attitudes and central topic content after the incident. And the detection of public opinion changes.In previous research on network public opinion, researchers mostly used qualitative analysis methods. For example, Kang Wei et al. used social network analysis (SNA) to study the problem of the spread of public opinion on the Internet [1].Some scholars also use the topic model to iterate the LDA model for many times to extract keywords and analyze the topic content [2].Jiang Jingui et al. used the Word2vec model to analyze the topic change in the relationship between topic feature words and similarity [3].

In the process of text semantic mining, the above methods are too simple and do not conduct in-depth analysis of text content. In response to the above problems, we propose

[1] Corresponding Author: Zhai Guangyu ; School of Economics Management, Lanzhou University of Technology, Lanzhou, Gansu 730050, China ; 2495696060@qq.com

a new topic model combining LDA and Doc2vec, namely the DBOW-LDA model, which is different from the previous analysis methods for network public opinion. The model is based on analyzing the topics on the entire corpus, combined with DBOW-LDA for topic detection on sub-data sets to obtain topic information. DBOW-LDA is an improved algorithm proposed by introducing LDA on the basis of DBOW.

2. Related Theory

2.1 Latent Dirichlet Allocation

LDA[4]is an unsupervised probabilistic topic model that can be used in large databases. The model has high interpretability and the algorithm runs relatively fast. Therefore, the LDA model is used to extract global topics.The calculation process of the LDA model can be divided into two steps: the first step is the generation process of the mth document, each word is obtained according to the probability distribution, and n words form the document m; the second step is to represent the mth document. The generation process of generating the nth word in m documents is to first select a topic from K topics [5], and then use this topic to generate the nth word. The model is shown in Figure 1.

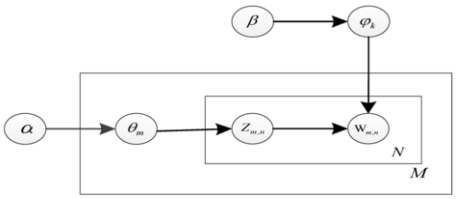

Figure 1. LDA Probabilistic Graphical Model

In LDA, each word is considered as discrete data extracted from a bag of words, each document is composed of a collection of words $w = \{w_1, w_2, ..., w_n\}$, and the corpus is composed of a series of documents $D = \{d_1, d_2, ..., d_n\}$. The joint probability distribution of LDA is expressed as

$$\rho(\theta, z, w \mid \alpha, \beta) = \rho(\theta \mid \alpha) * \prod_{n-1}^{N} \rho(w_n \mid z_n, \beta) \tag{1}$$

w represents the observed variable, θ and z represents the latent variable, α and β the sum is obtained by the EM algorithm.

2.2 Doc2vec

Doc2vec was proposed by Tomas Mikolov in 2014 based on the idea of the Word2vec model [6]. The word vector model uses the idea of neural networks to train sentences. At the same time, the model generates a corresponding word vector for each word. The model assumes that the occurrence of each word is only related to a specific number of words before it, focusing on the order combination between words. The basic idea of Word2vec is to predict the probability of the current word according to the context words. In this framework, each word is mapped into a vector, and the mapped word vector is used as the input matrix of W, with columns specified by the word indices in the phrase. These word vectors are used to predict the next word in the sentence. Doc2vec have two different representation models: distributed bag of words

(DBOW) and distributed memory (DM), which correspond to skip-gram and CBOW in Word2vec, respectively. The difference between the two models is that the PV-DM model predicts the target word based on the model in the given context.The model is shown in Figure 2.

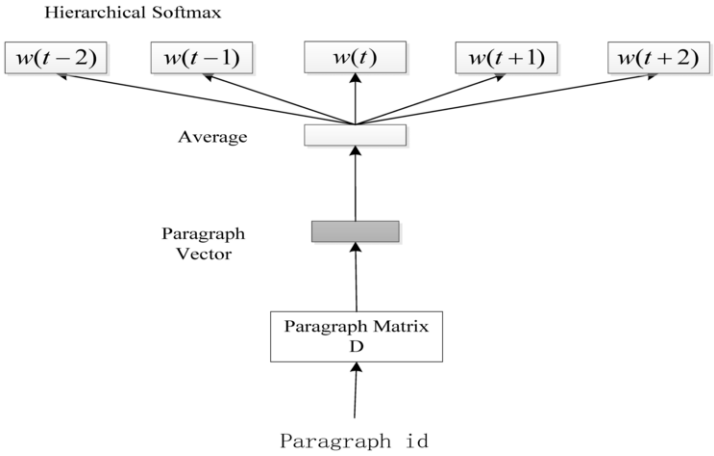

Figure 2. Training Doc2vec vector model with DBOW technique

In this framework, assuming that there are N texts in a document, each text is mapped to an independent vector of a specified dimension as a column of matrix D, and one of the columns of matrix W is the vector to which context words are mapped, corresponding to Word Matrix W in Figure 2. the paragraph vector Paragraph Vector in D and the word vector of W represent the Word Matrix W for summing or splicing to predict the probability of the target word, this process corresponds to the average operation in the middle layer in Figure 2.

3. Text Semantic Representation Based on LDA and Doc2vec

The document vector and topic vector generated by LDA and Doc2vec are mapped to the same vector space and then the similarity is calculated to complete the semantic representation of the text. The defect of this method is that the document vector and the topic vector belong to different dimensions, and direct mapping will cause Topic mapping is not accurate. In view of this, we proposes a new DBOW-LDA model, which incorporates the global topic information obtained by LDA into the sentence vector DBOW.

The principle of the model shown in Figure 3 is: in the input layer, each text is mapped into a column of text vectors with specified dimensions, the matrix D represents all paragraph vector matrices, the matrix M represents the vector matrix of all words in the text, and M is related to the topic distribution of the document .The product of is expressed as the topic vector of the document, and the topic distribution θ_{id} is obtained by LDA. The average value of the text vector and topic vector in the connection layer is used to predict randomly sampled words. The addition of LDA makes the Paragraph Vector model, which originally only has local text semantic information, have global semantic information.

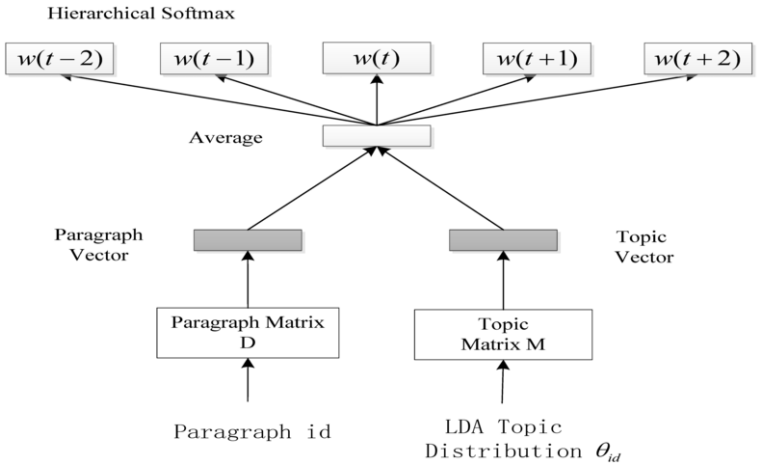

Figure 3. DBOW-LDA model

The DBOW-LDA model predicts words according to the maximum likelihood function, and its objective function is expressed as

$$F = \sum_{w \in corpus} \log P(\tilde{w} \mid \eta_w, \lambda_w) \tag{1}$$

The average of the text vector and topic vector in the connection layer is expressed as

$$z_w = \frac{\eta_w + \lambda_w}{2} \tag{2}$$

After the output layer is Huffman encoded, it is obtained by Hierarchical Softmax:

$$F_{\tilde{w}}^{j} = c_{\tilde{w}}^{j} \log \sigma(x_w^T \gamma_{\tilde{w}}^{j})(1 - c_{\tilde{w}}^{j}) \log(1 - \sigma(x_w^T \gamma_{\tilde{w}}^{j})) \tag{3}$$

The algorithm steps are as following:

Input: corpus Doc, sampling window p, iteration number iter, learning rate

Output: Sentence vector values for each text

Step 1:Use the LDA model to get the topic distribution θ_i *of each text;*

Step 2: Randomly initialize text d_i sentence vector parameters: word vector v, text vector η and matrix M ;

Step 3: For words $\tilde{w}$, the average value z_w is calculated by formula $z_w = \frac{\eta_w + \lambda_w}{2}$, and the gradient and parameters $\gamma_{\tilde{w}}^{j}$ of the topic matrix are consistent with the calculation method of the vector matrix. Among them, the vector gradient $e := e + \varepsilon \frac{dL_w^j}{dz_w^j}$,

Step 4: Update the text vector $\eta = \eta + e$;

*Step 5:*Update the subject matrix $M = M + e_M$, e、 e_M both are errors;

Step 6: Return to step 2 to train the next word;

Step 7: Repeat the above steps in sequence iter times.

The new framework mainly supplements the topic semantic information of the input layer of Doc2vec by introducing the topic distribution of LDA. The main idea of the

improved algorithm comes from BTPV-DBOW, and the BTM topic sampling in it is replaced by the Gibbs sampling of the LDA model. The rest of the topic inference process is consistent with BTPV-DBOW. In the middle layer of the sentence vector, the topic distribution obtained by LDA is vectorized and then averaged with the text vector in DBOW, and finally the sentence vector of the given text containing the topic semantic information is output. Express. In the experimental process, the global characteristics of LDA to extract topics and the local characteristics of paragraphs and contexts are used to identify the topic words in the network public opinion. The corpus is quickly expressed in vector form with the training model. Words to describe the subject information of an event.

4. Experiment and Analysis

Compared with the summary data in HowNet, the text of popular comments on Weibo is more complex, with a larger number of irrelevant words and higher requirements for text representation. Therefore, the comment information is used as the data set. First, the LDA model is processed to obtain the document topic, and then the sub-data set is represented by the sentence vector through the DBOW-LDA model proposed in the previous section to obtain the local topic. Finally, the accuracy, recall and F1 are used to evaluate the performance of the experimental method.

4.1 Experimental program

The data set is derived from the comment information about the popular events of "Didi Scandal" crawled on Weibo. It is crawled with the request library. After data collection, the data is divided into all data sets and sub-data sets. The sub-data set is the comment information of each day divided by the number of days as the time node . The two parts of the data were processed by word segmentation and stop word removal, and there were 78,233 pieces of processed data. Then, the corpus formed by all the data is used for topic modeling through the LDA model.Set up the LDA model $\alpha = 50/K$, $\beta = 0.01$ .The value of Doc2vec are set to be consistent with those in Table 1. Among them, among the hyperparameters , the value of the number of topics N in LDA is subjectively selected, and finally Get the topic distribution for each text.

Table 1. Doc2vec parameter settings

Parameter	Value
Size	100
Window	10
min_count	2
Workers	2
Dm	1

Compare the experimental parameters of the DBOW-LDA model proposed in this paper with the Paragraph vector and Word2vec models. The experimental comparison is mainly in the accuracy, recall and F value of text clustering under each model, and the sampling window p=20 in the Paragraph Vector model is set, the number of iterations =20, the learning rate ε=0.025 , the K-means clustering algorithm is used in the experiment to verify the validity of the model, K-means belongs to the plane division method, the algorithm is theoretically reliable, fast, easy to implement, and has low

dependence on data. It is also suitable for clustering analysis of various data such as text and image features [7]. Since the F value plays a reconciling role in the precision and recall rate, we focus on the analysis of the F value in the experimental results.

4.2 Experimental results

After data preprocess, LDA model is used to detect document topics in the entire data set. The LDA model represents each document as a random combination of words of potential topics, where each topic is composed of words with topic distribution probability. Select N=8, extract six words from each topic, and get eight topics as shown in Table 2. The topic obtained by the LDA model is represented by a cluster of related words and the probability of the word appearing. In order to further analyze the public opinion information, we define the topic label through manual annotation. It can be seen from the document theme that the crawled comment information mainly focuses on the platform, suspect information, travel software, social supervision, women's safety, etc.

Table 2. Document topics obtained by the LDA model

Topic	Keywords	Hashtags
Topic#0	Airline-stewardess、passenger、complaint、murder case、intention、car-hailing	Event description
Topic#1	Check、information、data、manage、company、enterprise	company information
Topic#2	license plate、daytime、Yueqing、suspects、Shady、twice	license plate shady
Topic#3	Platform、travel、Shouqi、Dida、Caocao、shenzhou	travel software
Topic#4	Supervision、filing、National、audit、killed、department	social regulation
Topic#5	Strength、punish、rectification、take down、software、manage	Rectify the industry
Topic#6	Safety、female、software、contact、position、report	female safety
Topic#7	Platform、criminal record、Li Min、the day before、complaint、mistake	similar cases

Document topics obtained from LDA can see what the public is concerned about after the whole event, but it is not possible to draw out how the event's focus has changed over time. Therefore, the topic evolution is obtained by expressing the similarity of the corresponding keywords on the sub-data of each day by the newly constructed model DBOW-LDA.

On the basis of the document topic obtained by LDA, we use DBOW-LDA to express the similarity of the data on the sub-data set. The sub-data set is divided by the number of days as the time node. In this experiment, the keywords in the event are used. as the central word to express related words.As shown in Table 3.The processed data set is trained by DBOW-LDA, and takes "DiDi" as the central word, and the corresponding correlation selects the first 10 words according to their occurrence probability. In the document topic obtained by LDA, the appearance of the word "rectification" can be seen. In order to further represent the topic in the stage, we use this word as the central word for similarity representation, and get the topic shown on the right.The similarity results of the document topics obtained by the above LDA and DBOW-LDA indicate that the hot topics that appear after the whole event and the evolution of the topic content over time can be detected.

Table 3. Some DBOW-LDA similarity results

Central Word	DiDi	Cosine	Rectification	Cosine Distance
Related Words	hitchhiker	0.78736	online	0.65253
	girls	0.76233	examination	0.62535
	killed	0.73843	learn a lesson	0.52534
	Yueqing	0.68635	Ministry of	0.51988
	car-hailing	0.65234	DiDi	0.49263
	rectification	0.57765	remediation	0.48762
	apologize	0.54068	stationed	0.45266
	Zhao	0.53945	measure	0.42232
	criminal record	0.52311	cancel	0.42112
	platform	0.42341	department	0.35551

4.3 Performance comparison

The DBOW-LDA algorithm proposed in this paper mainly introduces the topic distribution information of LDA into the Paragraph Vector model to improve the semantic discrimination of sentence vectors in similarity representation [8];On the basis of the sentence vector model, DBOW-LDA not only changes the distributed representation model, but also performs the weighting of the average value of the maximum likelihood function.In the network public opinion analysis based on topic model and sentence vector, the LDA model is used to represent the document topic. The parameter analysis is divided into two parts, one is the analysis of the number of topics in the LDA modeling,the other part is the experimental evaluation of the vector dimension, sliding window size, etc. in DBOW-LDA [9]. In order to verify the effectiveness of the DBOW-LDA algorithm proposed in this chapter, a comparison experiment is carried out with Word2vec, WT-LDA and LDA+Word2vec algorithms.

This paper selects K-Means, takes the text topic matrix as the input of K-Means, minimizes the distance between the cluster center and the corresponding document words, and finally obtains the clustering result. The evaluation of the clustering algorithm uses the precision rate, recall rate and F value to comprehensively evaluate.

1). Number of topics

When conducting network public opinion analysis, we first use the LDA model to represent document topics. From the perspective of topic labeling of words, too many topics will lead to too detailed distinction of each category between topics, resulting in words with low semantic similarity values appearing under different topics, and too few topics will lead to similar polysemy words appears under the same theme, resulting in the theme resolution is too low. Therefore, the range of the number of topics is selected as {5,40}, and the interval is 5. According to Figure 4, it is most suitable when the number of topics is 10. Accuracy for this number of topics is the highest.

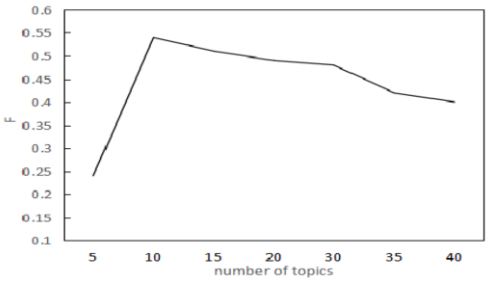

Figure 4. Parameter estimation for number of topics

2). DBOW-LDA vector window size

In the training process of the DBOW-LDA model, the size of the sampling window P value will affect the semantic training results of the current word. Therefore, a quantitative evaluation of different sliding window sizes in the algorithm is carried out. It can be seen from Figure 5 that the overall F value of the size of the vector window increases, but when P=20, the increase of the F value gradually becomes stable. It can be seen that as the window size grows to a certain length, continue Increasing the window size has no significant effect on the improvement of the F value, and the increase of the vector window size will lead to an increase in the complexity of the algorithm. Therefore, this paper selects P=20 as the window length.

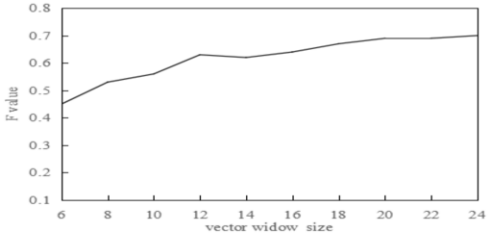

Figure 5. Estimation of vector window size

3). DBOW-LDA vector dimension

The size of the word vector dimension indicates that it affects the judgment of word similarity. In theory, the higher the vector dimension, the more accurate the model training result, that is, the better the word similarity expression result. However, in practice, if the vector dimension is too high, the training time will increase and the algorithm complexity will also increase. Therefore, we use the F value to evaluate the value of the vector dimension. It can be seen from Figure 6 that when the vector dimension is 200 , the F value is the highest, and then the increase of the dimension size F value gradually tends to be flat. Therefore, the size of the vector dimension is selected as 200 for clustering experimental evaluation.

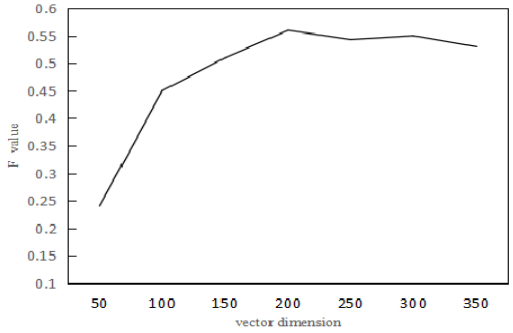

Figure 6. Parameter estimation for vector dimension size

4). Performance comparison of different algorithms

Through the above parameter estimation of the number of topics, the size of the vector window and the size of the vector dimension, the values of each algorithm in the DBOW-LDA algorithm are determined. performance of the algorithm.In order to further verify the comprehensive effect of DBOW-LDA model on text clustering, we selected Word2vec, WT-LDA, LDA+Word2vec and the new model for clustering comparison.

LDA: This model is based on the bag-of-words model for topic division, and words with the same topic are classified into the same class.

WT-LDA [10]: This method adds word label information on the basis of traditional LDA, and each word has the highest probability of belonging to the topic category and words with the same topic are classified into the same category.

LDA+Word2vec [11]: This method calculates the similarity of the topic matrix and document matrix obtained by the LDA model and Word2vec respectively, and then uses the K-means algorithm to perform text clustering on the processed comment information.

DBOW-LDA: This method is the algorithm proposed in this chapter. The method first uses LDA to obtain topic distribution, then averages the topic vector and text vector, and then obtains the sentence vector representation of the text, and calculates the text according to the K-Means algorithm.

Table 4. Comparison of results of different clustering methods

Clustering model	Accuracy	Recall	F1
Word2vec	0.489	0.562	0.523
WT-LDA	0.562	0.551	0.556
Doc2vec	0.532	0.512	0.522
LDA+Word2vec	0.664	0.612	0.637
DBOW-LDA	0.652	0.631	0.641

It can be seen from Table 4 that for text clustering, the DBOW-LDA method has obvious advantages compared with other methods such as LDA, WT-LDA, LDA-K and LDA+Word2vec, and the methods proposed in this paper are improved respectively. This is because the topic model obtained by combining the global semantic information of LDA with the local semantic information of Doc2vec has better clustering effect on comment information and more comprehensive semantic representation. Compared with Word2vec and Doc2vec, the topic model algorithm with LDA has better performance, and the text clustering effect of the Doc2vec model is better than that of Word2vec, because Doc2vec has the addition of paragraph vectors, which is more accurate for the contextual representation of words. In this chapter, the DBOW-LDA proposed by combining Doc2vec and LDA has a higher F value in the application of microblog comment information with a larger corpus. This is because the topic vector and text vector are connected, so that the new model has more for comprehensive semantic information [12].

5. Conclusions

This paper first analyzes various methods for online public opinion in text semantic mining. Based on the analysis of network public opinion, a new DBOW-LDA algorithm is proposed, and a topic evolution model of network public opinion based on topic model and sentence vector is constructed. The model is based on the topic analysis of the entire corpus by the LDA algorithm, combined with DBOW-LDA for topic detection on the word similarity representation on the sub-dataset. Then we crawled the comment information of the Didi scandal in Weibo, and introduced the setting method of various parameters in the experimental process in detail. The experimental results are analyzed in two aspects. First, the text topics obtained by LDA and the phase similarity results obtained by the DBOW-LDA model are used for qualitative analysis, and then the relevant parameters are quantified by using the precision rate, recall rate and F value. Finally, by comparing the performance of the clustering results with the original model

and other text representation methods, it is verified that the model proposed in this chapter has better performance in text semantic mining.

Acknowdgements

This work was financially supported by the National Natural Science Foundation of China (Grant No.71861026).

References

[1] Kang Wei. Social Network Structure Measurement and Analysis of Public Opinion Dissemination in Emergencies: An Empirical Study Based on "11.16 School Bus Accident" [J]. China Soft Science, 2018(07): 174-183.(in Chinese)

[2] Yu Bengong, Zhang Weichun, Wang Longfei. Topic detection and evolution analysis based on improved OLDA model [J]. Journal of Intelligence, 2017(02): 106-111.

[3] Jiang Jingui, Yan Siqi. Research on Weibo Public Opinion Evolution Based on Interaction between Theme and Emotion: Taking "Red, Yellow and Blue Child Abuse" as an Example [J]. Journal of Intelligence, 2018, 37(12): 118-123 .

[4] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learningresearch, 2003, 3(Jan): 993-1022.

[5] Mahdizadeh H, Biemans H, Mulder M. Determining factors of the use of e-learning environments by university teachers[J]. Computers & Education, 2008, 51(1): 142-154.

[6] Le Q V, Mikolov T. Distributed Representations of Sentences and Documents[J]. Computation and Language, 2014. (45): 234-241.

[7] Jin Jianguo. Overview of Clustering Methods [J]. Computer Science, 2014, 41(b11): 288−293.

[8] Rao Yuhe, Ling Zhihao. A short text clustering method combining topic model and paragraph vector [J]. Journal of East China University of Science and Technology (Natural Science Edition), 2020, 4(5): 23-31.

[9] Li Siyu. Research on short text semantic mining based on topic model and word vector [D]. Taiyuan: Taiyuan University of Technology, 2018.

[10] Chen L, Wang Y, Yu Q, et al. WT-LDA: User Tagging Augmented LDA for Web Service Clustering[C]. International conference on service oriented computing, 2013(6): 162-176.

[11] Li C, Lu Y, Wu J, et al. LDA Meets Word2Vec: A Novel Model for Academic Abstract Clustering[C]// Companion of the Web Conference 2018. 2018(5): 34-41.

[12] Peng Huaijin. Research on text representation model based on LDA and latent feature vector [D]. Beijing: Beijing University of Posts and Telecommunications, 2019.