A TOPIC MODEL BASED ON WEIGHTED WORD CO-OCCURRENCE MATRIX AND USER TOPIC RELATIONSHIPS

Ziqi Xu[1,2], Bo Cheng[1,2], Kang Yang[3], Lili Zhong[3], Yan Tang[3]

[1] Shenzhen Audencia Business School, WeBank Institute of Fintech, Shenzhen University, Guangdong, China
[2] State Key Laboratory Of Networking And Switching Technology, Beijing University Of Posts And Telecommunication, Beijing, China
[3]Ping An Bank Co., Ltd.

ABSTRACT

Various industries have widespread adopted the intelligent customer service, thus how to understand customer intent more accurately and extract key information has become a current research hotspot. However, the features that customer service dialog texts are short length, specialization and sparse lead to the poor performance of traditional topic extraction. Based on the above background and characteristics, this paper proposes a topic model WCMUT-HDP, which is based on a weighted word co-occurrence matrix and user topic relations. In the WCMUT-HDP model, this paper introduces a semantically weighted word co-occurrence matrix to mine the statistical and semantic features of customer service texts and optimize the effect of clustering. For the structure of customer service dialogues, this paper introduces temporal and author attributes of customer service dialog into the topic recognition of customer service texts. This method helps us to accurately extracts the user's intention. WCMUT-HDP is based on the Dirichlet process, and does not need to specify the number of topics in advance, which saving the time overhead of parameter experiments and evaluation. In the end of this paper, the experimental results show that the WCMUT-HDP model can effectively identify the topics of customer service conversations, and the extracted topics can accurately reflect the user's conversational intent.

KEYWORDS

Customer Service Text, Word Co-occurrence Matrix, Hierarchical Dirichlet Process, Topic Model

1. INTRODUCTION

With the rapid development of Internet technology, the information age has arrived. Companies can improve their service quality by listening to users' demands and providing timely feedback through customer service systems. The extraction of dialogue topics for customer service texts has become a hot topic of research in the field of natural language processing currently. Customer service text is a typical interactive short text, which is short in length and sparse in semantic features. Traditional topic extraction models are based on bag-of-words models, which are more effective in extracting topics on long texts. It does not perform well for texts with highly sparse features such as customer service dialogues.

David C. Wyld et al. (Eds): SIPM, ITCA, CMIT, FCST, CoNeCo, SAIM, ICITE, ACSIT, SNLP -2023
pp. 157-167, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130917

In order to extract more text information, some researchers have introduced word co-occurrence relations into topic recognition, mining the statistical features of words by counting the co-occurrence relations of words. However, the customer service dialogues are short in length and the co-occurrence relationships of words are limited, which makes insufficient contribution to topic clustering. At the same time, customer service texts contain several additional attributes, such as time attributes and author attributes. These attributes can be introduced into the topic extraction of the text as well.

Based on the above background, this paper proposes a topic extraction model WCMUT-HDP based on a weighted word co-occurrence matrix and user topic relations. This model enriches the short text information by introducing a weighted word co-occurrence matrix to extract more features from the dialogue text. The variation in user preferences for topics is continuous over time. By introducing the topic distribution of past users the current topics can be better extracted.

2. RELATED WORK

Customer service dialogues, as a typical short interactive text, are characterized by short length, spoken language, and sparse features. Traditional topic extraction models cannot accurately extract topics from such texts. For this type of text, researchers have proposed a series of topic models to explore its potential information, enrich the semantics of the text, and optimize the effect of topic extraction.

The co-occurrence relationship of words can reflect the semantic information of words to a certain extent. Some scholars have introduced word co-occurrence matrices into topic recognition.

Xu [1] mined sub-topics related to hot topics through word co-occurrence relations and subdivided the topic extraction results. Zhang [2] used the co-occurrence relations of feature words in each time window to construct a co-word network and used the LPA algorithm to find communities in the co-word network through community discovery techniques, where each community represented a sub-topic. However, the above methods are only based on statistical features to describe word co-occurrence relationships, and more semantic information between words can be mined from customer service dialogue texts.

Li[3] adopt word co-occurrence network in social network topics analysis. Based on the word co-occurrence network, Lidis covered the relevance of social network topics and quickly found hot topics. Zhang[4] added semantic information to the co-occurrence network, which can more accurately and quickly discover the clustering of topics.

In addition to traditional text topic extraction, topic extraction of web texts is also a hot research topic [5]. Wang [6] proposed a multi-attribute latent Dirichlet distribution model (MA-LDA). The model takes into account additional attributes of micro blog texts, such as time and hash tag attributes.MA-LDA can filter the collection of popular topics based on these attributes. Zhang [7] et al. mined the citation relations of academic papers and link relations of web pages and proposed a holistic topic model to capture semantic information.

In this paper, word co-occurrence relationships and user and time attributes in customer service texts are introduced into topic recognition to mine more information based on the structure of customer service texts, and extract the topics of customer service texts more accurately by mining their potential information.

3. A TOPIC MODEL BASED ON WEIGHTED WORD CO-OCCURRENCE MATRIX AND USER TOPIC RELATIONSHIPS

WCMUT-HDP is a topic extraction model based on a weighted word co-occurrence matrix and user-topic relations, which is based on the HDP model and introduces a semantically weighted word co-occurrence matrix in the clustering process. Users are rich in personal colours and have obvious personal preferences. WCMUT-HDP can better understand the focus of users through the relationship between users and topics, and obtain better topic extraction results.

3.1. The Weighted Word Co-Occurrence Matrix

The co-occurrence matrix is an important method of representing distributed text. A matrix is formed by counting the number of simultaneous occurrences of words in a text, with each element of the matrix representing the number of times two words appear together in the same window.

As semantically similar words often belong to the same topic, co-occurrence matrices can also play an important role in topic models. In this paper, we introduce a weighted semantic similarity to the traditional word co-occurren*ce matrix, so that the constructed co-occurrence matrix contains not only the location but also the semant*ic information of the words, thus portraying the underlying information of the text more accurately. A dialogue in the corpus is shown in Table 1:

Table 1. Dialog Example

No.	Example sentence
1	Hello, what kind of business do you need to handle?
2	I've got my credit card and I'd like to know how to use it.
3	You will need to activate your credit card before you can use it. Simply call the customer service number on the back and listen to the voice prompts.

After pre-processing the above example sentences, the verbs and nouns were extracted to construct the word co-occurrence matrix M_{co}. The specified window length was the length of the example sentence in which the words were located. The co-occurrence matrix was obtained as shown in Table 2.

Table 2. The co-occurrence matrix

	Business	Credit Card	Use	Activation	Customer Service	Phone	Voice
Business	0	0	0	0	0	0	0
Credit	0	0	2	1	1	1	1
Use	0	2	0	1	1	1	1
Activation	0	1	1	0	1	1	1
Customer Service	0	1	1	1	0	1	1
Phone	0	1	1	1	1	0	1
Voice	0	1	1	1	1	1	0

The list of words that co-occur with word w_i can be seen in the matrix M_{co} to characterise the meaning of word w_i jointly to a large extent, but only in terms of the statistical features that reflect the relationship between the words. In this paper, the semantic similarity of words is also

introduced into the co-occurrence matrix. Let the word vectors corresponding to the two words be A and B. The remaining string similarity is

$$similarity(A,B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

In this case, the A_i and B_i represent each component of vectors A and B respectively. The cosine similarity between words in the dictionary is calculated to obtain the word similarity matrix M_{sim}. The obtained similarity matrix M_{sim} is weighted into the word co-occurrence matrix M_{co}. The number of co-occurrence of words w_i and w_j is $CO_{i,j}$, the similarity is $Sim_{i,j}$, and the weighted word co-occurrence relationship is $M_{i,j}$. The formula is as follows:

$$M_{i,j} = \left(1 + CO_{i,j}\right) \times Sim_{i,j}, (i \neq j) \quad (2)$$

In this case, adding the number one to the number of co-occurrence is to avoid the failure of similarity weights caused by words that have not co-occurred. The weighted word co-occurrence matrix WCM (weighted co-occurrence matrix) was obtained according to equation (2) as shown in Table 3.

Table 3. The weighted co-occurrence matrix

	Business	**Credit Card**	**Use**	**Activation**	**Customer Service**	**Phone**	**Voice**
Business	0	0.68	0.51	0.47	0.63	0.66	0.44
Credit	0.68	0	1.23	1.08	1.11	1.07	0.85
Use	0.51	1.23	0	1.33	0.85	0.82	0.92
Activation	0.47	1.08	1.33	0	1.17	0.84	1.01
Customer Service	0.63	1.11	0.85	1.17	0	1.36	0.9
Phone	0.66	1.07	0.82	0.84	1.36	0	1.2
Voice	0.44	0.85	0.92	1.01	0.9	1.2	0

This paper proposes a weighted word co-occurrence matrix that mines not only the statistical information of the dialogue text, but also the semantic information between words. For example, the words "credit card" and "phone" are closely related to the word "business".

By referring to the topics corresponding to the words most closely related to the current words, the semantic information can be more fully utilised and the accuracy and robustness of the topic model can be improved.

3.2. User Topic Relations

In the context of customer service conversations, where users are rich in personal colours and have distinct personal preferences, exploring the relationship between users and topics can help the topic model to better understand the focus of users' attention. This paper proposes a user-topic relationship model based on the author-topic relationship model [8].

Traditional topic models are usually modelled based on the statistical features of the text body, which add a layer of topics between the document and the words that make it up, representing the document-word distribution as a document-topic as well as a topic-word distribution, and usually do not consider the relationship between the author of the text and the topic. The author-topic

relationship refers to the author's preference or level of interest in a particular topic. Thus the same user tends to have some preferences on different topics, and different users have different preferences on different topics. The user-topic model proposed in this paper introduces author-topic relations into the implementation of a textual topic model for customer service conversations, taking into account the relationships between users, texts and topics, and by exploring the performance of users on different topics, the potential relationships between users and topics can be revealed.

In this paper, the topics of all conversations of each user are counted to form a user-topic distribution (UT), which is used as information about the user's historical topic preferences. In the training process of the topic model, the selection of topics corresponding to a certain word is acted upon to better identify the information about the user's preferred topics.

3.3. Joint Enhanced Topic Model

WCMUT-HDP introduces the semantic features from the weighted word co-occurrence matrix described above and the user-topic model into the topic model to form a jointly enhanced topic model. The model is based on the Hierarchical Dirichlet Process (HDP) model, a non-parametric extension of the traditional LDA model, which allows for adaptive selection of the number of topics.

The Chinese restaurant franchise (CRF) is usually used to simulate the HDP construction process. In the training process of the topic model, a corresponding topic is selected for each word.

WCMUT-HDP introduces a weighted word co-occurrence matrix and a user-topic relationship into the HDP construction process to optimize the performance of the topic model in short texts, as shown in the following graphical model of WCMUT-HDP:

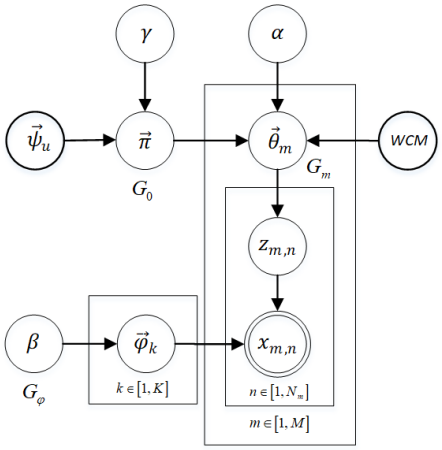

Figure 1. The WCMUT-HDP Diagram Model

WCM is the weighted word co-occurrence matrix, which acts on clustering together with the scale parameter α; $\vec{\psi}_u$ is the user topic preference distribution and the scale parameter γ acts on topic selection. In this paper, a Gibbs sampling algorithm based on the CRF model is used to sample the WCMUT-HDP model for training. The main idea is to keep other variables constant and sample a particular variable in the model. In the construction of the CRF, firstly, tables are assigned to the words in each document; then topics are assigned to each table. Denote t_{ji} as the cluster corresponding to the word i in the document j, and k_{jt} as the topic corresponding to the

cluster t in the document j. Gibbs sampling is the sampling of each cluster t and topic k in the CRF.

(1) The first step is to sample the clusters t.
The process of constructing a cluster is similar to assigning a table to each customer; the more people seated at a table, the greater the probability that a customer will be assigned to that table. Denote the probability that customer x_{ji} is assigned to table ψ_{jt} is as follows:

$$p = \frac{n_{jt}}{i - 1 + \alpha} \tag{6}$$

As Eq.6, n_{jt} is the number of customers at the table t in the restaurant j. There is also a certain probability that a customer will be assigned to a new table, and the probability of being assigned to a new table ψ_{jt}^{new} is:

$$p = \frac{\alpha}{i - 1 + \alpha} \tag{7}$$

Therefore, the conditional probability that customer x_{ji} is assigned to the table t is:

$$p\left(t_{ji} = t \middle| t^{-ji}, k\right) \propto \begin{cases} n_{jt}^{-ji} f_{k_{jt}}^{-x_{ji}}\left(x_{ji}\right), & \text{table t has been used} \\ \alpha p\left(x_{ji} \middle| t^{-ji}, t_{ji} = t^{new}, k\right), & t = t^{new} \end{cases} \tag{8}$$

As Eq.8, n_{jt}^{-ji} denotes the number of customers on the table t at the restaurant j that does not include the customer i. $f_{k_{jt}}^{-x_{ji}}\left(x_{ji}\right)$ denotes the conditional probability of the observation x_{ji}.

According to the weighted word co-occurrence matrix WCM, words with similar locations and semantic meaning in the same sentence will have a higher probability of belonging to the same topic. By arranging the data in the columns of the matrix in descending order, the current word has a certain probability of being assigned to the same topic as the word with the highest similarity. Let the word with the highest similarity to the word x_{ji} be x_{ji}^{sim}, and the Eq.6 and Eq.7 are modified to the following form:

$$p = \begin{cases} \dfrac{n_{jt^{sim}}}{i - 1 + \alpha} \cdot 2, & t = t_{ji}^{sim} \\ \dfrac{n_{jt}}{i - 1 + \alpha} \cdot \dfrac{i - 1 - 2 \times n_{jt^{sim}}}{i - 1 - n_{jt^{sim}}}, & t\ is\ other\ table\ that\ has\ been\ used \\ \dfrac{\alpha}{i - 1 + \alpha}, & t = t^{new} \end{cases} \tag{9}$$

In this equation, $n_{jt^{sim}}$ is the number of customers at the table where customer x_{ji}^{sim} is seated. Equation (9) doubles the probability of a customer being assigned to table t_{ji}^{sim}, makes a year-on-year reduction in the probability of being assigned to another table, and leaves the probability of being assigned to a new table unchanged. Following such an allocation can over make customer x_{ji} more likely to choose the table where a customer with a high degree of similarity to him is located, and his probability of choosing all tables still sums to one.

(2) The Next step is to sample the theme k

After all table assignments have been made, the tables need to be assigned dishes. If a customer is seated at a table that is already occupied, he will share the dish chosen by the first person; if a customer chooses a new table, he will need to be assigned a dish based on the dishes already chosen. k_{jt} is similar to t_{ji} in that it is also proportional to the number of tables at which dish k is chosen, and the probability that customer x_{ji} chooses dish ϕ_k, which has already been chosen, is:

$$p = \frac{m_k}{\sum_k m_k + \gamma} \tag{10}$$

As Eq.10, m_k is the number of tables offering dish k and $\sum_k m_k$ is the number of all tables offered. Customers also have a certain probability of choosing a new dish, and the probability of choosing a new dish ϕ_k^{new} is

$$p = \frac{\gamma}{\sum_k m_k + \gamma} \tag{11}$$

The conditional probability k_{jt} of choosing dish k at table t is

$$p\left(k_{jt} = k \middle| t, k^{-jt}\right) \propto \begin{cases} m_k^{-jt} f_k^{-x_{jt}}\left(x_{jt}\right), k \text{ has been selected before} \\ \gamma f_{k^{new}}^{-x_{jt}}\left(x_{jt}\right), \quad k = k^{new} \end{cases} \tag{12}$$

m_k^{-jt} is the number of tables that select dish k excluding table t. In customer service conversations, there is a very close connection between the text and the user, and mining out the user's preferences helps to identify the topic of the text. By analysing information about users' past topic preferences, a user topic focus curve can be constructed, and topics with the high focus can be assigned in priority for the current conversation text. When assigning topics to dishes at the table, i.e. to clusters, there is also a certain probability that topics with an increasing trend in topic preference will be selected. The topic that is of most interest to the user at the moment is $\mathrm{k^h}$, which has either already been used at the table or may have never been assigned. If $\mathrm{k^h}$ has already been used, equations (10) and (11) are improved to the following form:

$$p = \begin{cases} \dfrac{m_{\mathrm{k^h}}}{\sum_k m_k + \gamma} \cdot 2, & \mathrm{k} = \mathrm{k^h} \\ \dfrac{m_k}{\sum_k m_k + \gamma} \cdot \dfrac{\sum_k m_k - 2 \times m_{\mathrm{k^h}}}{\sum_k m_k^{-\mathrm{k^h}}}, & others \\ \dfrac{\gamma}{\sum_k m_k + \gamma}, & k = k^{new} \end{cases} \tag{13}$$

As Eq.13, $\sum_k m_k^{-\mathrm{k^h}}$ is the total number of tables allocated that are not dishes $\mathrm{k^h}$. Similar to equation (13), equation (14) doubles the probability of allocation to $\mathrm{k^h}$ and reduces the probability of allocation to other dishes. If $\mathrm{k^h}$ has not been used, equations (10) and (11) are improved to the following form:

$$p = \begin{cases} \dfrac{m_k}{\sum_k m_k + \gamma}, & k \text{ is the dish that has been assigned} \\ \dfrac{\gamma}{\sum_k m_k + \gamma} \cdot \dfrac{1}{2}, & k^{new} = \mathrm{k^h} \\ \dfrac{\gamma}{\sum_k m_k + \gamma} \cdot \dfrac{1}{2}, & k^{new} \text{ is other dishes} \end{cases} \tag{14}$$

In this paper, the weighted word co-occurrence matrix and user topic relations are introduced into the CRF construction process. The use of the weighted word co-occurrence matrix influences the clustering process of words, making it easier to assign words with higher similarity to a category. Using user topic relations as a reference for selecting topics for clustering better reflects the information about users' preferences for topics. Such a training process not only mines the statistical and semantic features of the text but also incorporates the user's preference information for topics into the recognition of topics, thus enriching the semantic features of the text and better performing the topic extraction task.

4. EXPERIMENTS

4.1. The Dataset

This section uses the Chinese text classification dataset of t news to construct a short text corpus for testing, an example of which is shown in the following figure:


```
{
    "label": "114",
    "label_desc": "news_stock",
    "sentence": "Why continuous up to suspend the investigation and continuous down without, is it easier to
    endanger the stability of the securities market than a halt?",
    "keywords": "Suspension,Change of hands,Unusual volatility,Average daily change of hands,One-stop"
},
{
    "label": "104",
    "label_desc": "news_finance",
    "sentence": "05.09 Financial dinner: offline finance unreliable: Zhongrong Minxin rumored to be
    investigated",
    "keywords": "Roadheader,MBG,Lenovo Group,offline finance,vocational skills training,workers,RQFII"
},
{
    "label": "109",
    "label_desc": "news_tech",
    "sentence": "Achieving full coverage of financial supervision must be made real local financial
    supervision",
    "keywords": "regulatory coordination, financial regulatory framework, Financial Stability Development
    Committee of the State Council, finance, local finance"
}
```


Figure 2 Example of Dataset

In this paper, news data from the three categories of finance, stock, and technology in the dataset were extracted for comparison experiments. The content in the test dataset contains only news headlines, and the customer service dialogue texts studied in this paper are all of the short length and sparse text features, which meet the requirements of this section of the experiments. The experiments in this section randomly assigned each category of corpus to multiple authors to replace the user information in the customer service text data.

4.2. Compare Model Introduction

The following models have been selected as comparison models in this section:

(1) LDA: a widely used classical Bayesian parametric model that requires the number of topics to be manually specified. The optimal number of topics for the LDA model was determined to be three by parameter debugging in this section.
(2) K-means: An unsupervised clustering algorithm that classifies clusters by the similarity between data.

(3) HDP: an improved version of the Bayesian non-parametric model. In this section, a topic induction step is added to the HDP model to avoid overly granular extraction results due to semantic sparsity of the test corpus.

4.3. Results

In this section of the comparison experiment the dataset was entered into WCMUT-HDP and its comparison model separately and three metrics, Precision, Recall and F1-score, were used to evaluate the results. The evaluation results for each category and its macro-average are shown in Table 4.

Table 4 Comparison of classification results of various models

The model	Category	Precision	Recall	F1-score
LDA	Stock	0.64	0.49	0.55
	Finance	0.56	0.71	0.62
	Tech	0.68	0.66	0.67
	Macro avg	**0.63**	**0.62**	**0.62**
K-means	Stock	0.49	0.40	0.44
	Finance	0.37	0.73	0.49
	Tech	0.96	0.19	0.32
	Macro avg	**0.61**	**0.44**	**0.42**
HDP	Stock	0.49	0.66	0.56
	Finance	0.48	0.57	0.52
	Tech	0.79	0.35	0.49
	Macro avg	**0.59**	**0.53**	**0.52**
WCMUT-HDP	Stock	0.75	0.86	0.80
	Finance	0.79	0.26	0.39
	Tech	0.63	0.97	0.76
	Macro avg	**0.73**	**0.69**	**0.65**

A comparative presentation of the macro-average data in the table above using a bar chart shows the results as follows:

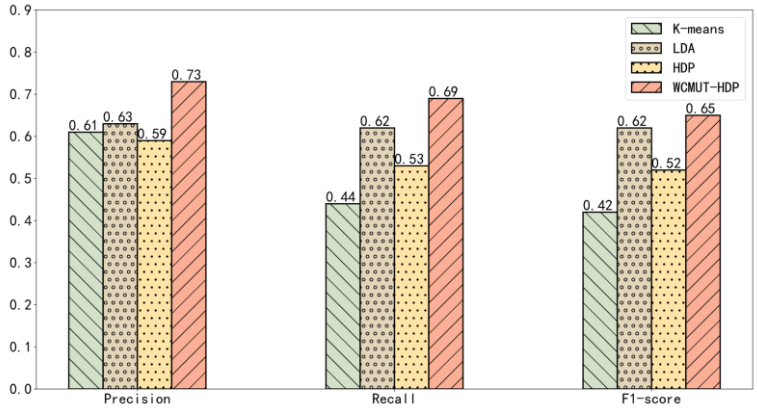

Figure 3 Comparison Chart of Accuracy, Recall, and F1-score of Each Model

As Fig.3, the WCMUT-HDP model proposed in this paper has improved in all metrics. Accuracy is improved by 10% compared to the LDA model, 14% compared to the HDP model, and recall is improved by 7% compared to LDA and 16% compared to HDP. The F1-score index also improved by 3% compared to the LDA model and by 13% compared to the HDP model.

The experimental results show that the WCMUT-HDP model improves in all metrics and can effectively extract topics accurately in short texts.

5. CONCLUSIONS

In this paper, we propose a topic model WCMUT-HDP based on a weighted word co-occurrence matrix and user topic relations to address the short length and sparse semantic features of current customer service dialogues. WCMUT-HDP introduces the similarity between word pairs through the word co-occurrence matrix for weighted summation, and the statistical information and semantic information of the text can be mined at the same time through the weighted word co-occurrence matrix. At the same time, the WCMUT-HDP model constructs the distribution of user-topic relationships by counting the distribution of users' conversation topics. The user-topic relationship is the user's preference distribution of the topic, which introduces the user preference information into the recognition of the text subject, which can more accurately reflect the user's dialogue intention. WCMUT-HDP is based on the Dirichlet process, which does not require specifying the number of topics in advance, saving the time overhead of parameter debugging and evaluation. Experimental results show that the WCMUT-HDP model can effectively identify the subject of customer service conversation, and the extracted topic can accurately reflect the user's conversation intention.

ACKNOWLEDGEMENTS

This work is supported by Swift Fund Fintech Funding.
Bo Cheng is the corresponding author of this paper.

REFERENCES

[1] Xu Guixian, Yu Ziheng, Wang Changzhi, et al. Research on topic discovery technology for Web news[J]. Neural Computing & Applications, 2020, 32(1):73-83.

[2] Zhang Peng , Li Bicheng, Yang Ruipeng. Research on the Topic Evolution of Microblog Based on BTM-LPA[C]//Proc of the International Conference on Computer Science and Technology (CST2016). Shenzhen, China, 2017: 860-875.
[3] LI Yaxing, WangZhaokai, FengXupeng, et al. Microblog Topic Discovery Based on Real-time Word Co-occurrence Network [J]. Journal of Computer Applications, 2016, 36(5):1302-1306
[4] ZhangXiaofei, ChenHanghang, Zhang Chunhua Research on microblog subject word extraction based on semantic concept and word co-occurrence[J]. Information Science,2021,39(01):142-147
[5] Nguyen T , Do P . CitationLDA++: an Extension of LDA for Discovering Topics in Document Network[J]. 2018.
[6] Wang Jing, Li Li, Tan Feng, et al. Detecting hotspot information using multi-attribute based topic model[J]. PloS One, 2015, 10(10): e0140539.
[7] Delvin Ce Zhang, Hady W. Lauw. Topic modeling on document networks with adjacent-encoder[C]//Proc of the AAAI Conference on Artificial Intelligence, 2020, 34(04): 6737-6745.
[8] Rosen-ZviM , Griffiths T L , Steyvers M , et al. The Author-Topic Model for Authors and Documents[J]. AUAI Press, 2004.