Journal of Computers Vol. 32 No. 4, 2021, pp. 1-12
doi:10.53106/199115992021083204001

Topic Analysis in LDA Based on Keywords Selection

Bing-Xin Du[1*], Guo-Ying Liu[2]

[1] Media and Communication School of Anyang Normal University, Anyang Henan, 455000, PROC
dubingxin@126.com

[2] Computer and Information Engineering School of Anyang Normal University, Anyang Henan, 455000, PROC
LiuGuoY1979ay@sohu.com

Received 19 October 2020; Revised 16 April 2021; Accepted 22 April 2021

Abstract. The Latent Dirichlet Allocation model in text analysis has weak generalization ability and poor interpretability of the topic words. In this paper, we address these issues using a topic analysis framework for Latent Dirichlet Allocation based on keyword selection. Our proposed solution extracts the keywords from scientific research articles and builds a keywords list according to filter rules. Then several words are selected in the abstracts of the articles based on the keywords list and the LDA model is used to analyze the topics of the selected words. To evaluate the performance of our proposed approach, journal articles in the field of educational technology are selected as data sources, and two types of comparative analysis are performed. Firstly, "verb", and "verb + noun" word selection strategies are adopted to conduct a comparative study from aspects including domain expert analysis, model perplexity, topic coherence measure, and inter-topic distance analysis. Secondly, Hierarchical Dirichlet Process, Correlated Topic Models, and LDA-Word2Vec models are used to conduct a study from model perplexity and predictive log-likelihood aspects. The experimental results confirm that the topic analysis based on the keywords selection method overperforms others in both types of comparisons.

Keywords: inter-topic distance, Latent Dirichlet Allocation (LDA), model perplexity, topic coherence measure, topic model

1 Introduction

As a topic model based on probability and statistics, Latent Dirichlet Allocation (LDA) [1] provides a theoretical framework for analyzing and representing the semantic structure of large-scale document collections. The LDA and its extended versions have been widely used in the fields of information retrieval, document classification, topic detection, and evolution analysis [2]. The basic idea behind the LDA model is to assume that the words in a document in the corpus are a random mixture of latent topics, and the topic proportions are document-specific and randomly drawn from a Dirichlet distribution. Using unsupervised learning, the model can construct "word-topic" and "topic-document" matrices representing the probability distribution of documents' topics. The topics are therefore the semantic representation of the document collections.

Performing topic analysis, the LDA model regards the topic distribution and the word distribution as two random variables satisfying the Dirichlet prior distribution, where the hyperparameters of the prior distribution are α, β. According to the documents, the model then uses Collapsed Gibbs Sampling (CGS) or Variational Expectation-Maximization (Variational EM) [3] to estimate the document-topic posterior distribution matrix, θ, and topic-word posterior distribution matrix, φ.

The LDA model is capable of modeling document topics, nevertheless, this model is attributed to weak generalization ability, and poor interpretability of the topic terms [4]. This is mainly because the

[*] Corresponding Author

reasoning process of the LDA model is random. Hence, the analysis results are highly related to the setting of the model parameters, common words with high-frequency (after removing stop words), and text characteristics in different fields (e.g., using the original LDA model to analyze social network comments). To address the above issues, here we propose an LDA topic analysis framework based on the keywords selection method for scientific research articles' topic detection.

The main contributions of this work are listed in the following:

(1) We propose a topic analysis model based on keyword selection. The proposed model improves the readability and interpretability of the text topic analysis.

(2) We verify the model efficiency through two different types of comparative studies, where we compare the performance of topic analysis of academic literature with different word selection schemes, and then compare its generalization ability with three typical existing topic models. The experimental results confirm that the topic analysis based on the keywords selection method we proposed, overperforms others in both types of comparisons.

The remainder of this paper is organized as follows. In Section 2, we provide a brief review of the existing related works of the topic model. Sections 3 and 4 describe the framework of our research and evaluate the performance of our method. Conclusions and future work are presented in Section 5.

2 Related Works

In many existing works, the objective is to improve the analysis effect of the LDA topic model from the model optimization and word selection design perspectives.

2.1 Optimization of LDA Model

The topics of the original LDA model are independent and identically distributed. In the other words, the appearance of one topic has nothing to do with other topics. This ideal state is not necessarily correct. To address this issue, Lafferty [5] proposes Correlate Topic Modal (CTM), where the correlations of the probability distributions between topics are modeled using Logistic Normal Distribution and replace the document-topic Dirichlet distribution.

Furthermore, model perplexity is often used to determine the optimal number of topics. Perplexity is a parameter that measures how well the model fits with the true probability distribution. In the LDA topic model, the true probability distribution of document-topic words is based on a priori hypothesis. Therefore, relying on perplexity to determine the number of topics remains strongly subjective. To address this issue, Yau [6] designs a topic mining scheme based on the Hierarchical Dirichlet Process (HDP) which is a non-parametric Bayesian model that utilizes the nested Chinese Restaurant Process (nCRP) to learn the topic distribution.

To perform topic analysis on the short texts of online comments (e.g., Weibo, Twitter), Yin et al. [7] propose a Dirichlet polynomial mixed distribution model, GSDMM, based on CGS sampling. Unlike the probabilistic mixture of multiple topics in a document in the LDA model, the topic distribution in the GSDMM model follows the Dirichlet prior distribution while each document is only generated by one topic. This modification is more suitable for the actual situation of short texts such as online reviews where there is a low probability of co-occurring words between sentences. The GSDMM abandons traditional text features such as TF-IDF and BoW and uses the Movie Group Process to cluster and automatically determine the number of topics (clusters). This method effectively solves the problems of high-dimensional sparse matrix and a large amount of computation in the feature representation of short text vector space. Different from Yin's work, Yan et al. [8] propose a Biterm Topic Model (BTM) to solve the problem of sparse high-dimensional features of short text. This model extracts co-words throughout the text corpus and then conducts topic analysis for the co-words.

With the continuous development of deep learning technology, neural networks, and deep learning have been extensively used to improve the performance of LDA. For instance, the pre-trained word vector models (such as Word2Vec) are combined with LDA for topic analysis [9]. As a result, words with similar semantics can be assigned to the same topic with greater probability. Further, Ting-ting Wang et al. [10] use LDA's topic-word matrix and word vector model to generate the T-WV matrix and adopt clustering to determine the number of topics. Das R. et al. [11] also use Word2Vec as the word feature representations for LDA analysis, so that the words with similar vector distances are most likely distributed in the same topic. Moreover, neural networks are used to vectorize sentences or documents to train a document set that is more relevant to the actual semantic needs [12]. Building a more accurate topic model based on the combination of deep learning and LDA is a significantly valuable research direction. Nevertheless, efficient application of this approach is the generation of word vectors and the fitting of neural network model parameters requires large well-annotated datasets. The lack of large-scale well-annotated datasets is the main limiting factor in the versatile development of such approaches.

2.2 LDA Word Selection Strategies

In addition to the optimization of the model, word selection strategies are used to improve the efficiency of the LDA model [13]. A reasonable word selection strategy effectively improves the readability and interpretability of LDA model analysis results. Irrelevant words are often extracted considering using either part-of-speech (POS) selection or domain terms construction. In terms of the POS selection, topic analysis is usually carried out with "noun" or "noun + verb" [14]. In addition to the POS selection methods, domain terms can be considered through manual or machine learning methods.

Construction of domain terminology is however a relatively complicated process. To analyze the evolution of the research topic of the knowledge system, Zhang Y. et al. [16] propose using a 3-step domain terms generation strategy including data cleaning, knowledge-based word merging, and rule-based word merging. In the whole process, 3956 domain terms are then constructed using multiple existing English corpus dictionaries and manual intervention of domain experts. Although the efficiency of topic analysis based on domain terminology is likely better than that of the word selection based on POS, construction of domain terminology is often a complex and subjective process. To the best of our knowledge in the absence of a complete term dictionary and term mapping tools, there is no viable alternative in the existing works. In this paper, we propose constructing a keywords table to design an LDA topic analysis framework based on keyword selection.

3 Research Design

Fig. 1 illustrates the research framework design of the LDA topic analysis of keywords selection. This research mainly consists of three stages: data acquisition, preprocessing model analysis, and evaluation of the results. The main task of the first stage is to establish a keywords table based on the acquired data. The second stage then performs two types of model analysis. The first one is different word selection strategies, and the second one is three representative model improvement methods including CTM, HDP, and LDA-Word2Vec. In the final stage, different comparison methods are designed for the two types of model analysis.

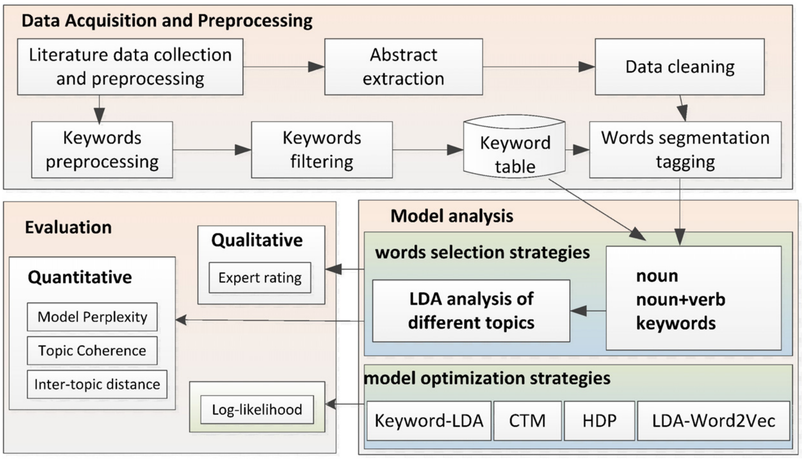


Fig. 1. The framework design for the LDA analysis is based on keyword selection

In word selection strategies, the keywords selection method is then compared with "noun" and "noun + verb" selection methods. The comparison is mainly conducted based on the expert rating, model perplexity, topic consistency, and inter-topic distance. The word selection strategy based on domain terminology used in the literature is not considered here. This is because in many fields, building a scientific and complete field terminology is still considered an open problem. In the comparative analysis of different model optimization strategies, the proposed method is compared with CTM [5], HDP [6], and LDA-Word2Vec [11]) based on their prediction log-likelihood.

3.1 Data Acquisition and Processing

To construct the keywords list and obtain the corpus for topic analysis, we use eight CSSCI journals in the field of education technology in China as the data sources. To increase the coverage of the keywords list, in the CNKI database the search time range is set to “2010-2019”. After excluding articles such as product introduction, conference notice, call for papers, conference summary, etc., a total of 12,633 valid articles are then obtained. We then export the reference materials in Endnote format and import them into SATI [18] for preliminary data conversion. After conversion, the keywords and abstracts of the papers are extracted and saved in Excel format for subsequent analysis.

After obtaining the keywords and the original data of the abstracts, we then perform the keywords screening stage. The keywords are first deduplicated, and their frequencies are calculated using the abstracts of the articles. In the LDA topic analysis, the number of co-occurrences of core words in the document has a significant impact on the results of the topic analysis. Note that without word selection the efficiency of LDA analysis is poor. This is because high-frequency co-occurring words are most likely just general words. Therefore, in performing keyword filtering, words with high keyword frequency are removed. Here we eliminate words with frequencies more than 1000 through experimental comparison. Furthermore, to improve the efficiency of LDA operations and reduce the sparsity of the corresponding feature matrix, we remove words with low frequency. Note that low-frequency words have a lower impact on the model analysis. Experimental analysis confirms that eliminating words with a frequency less than 10 has an ignorable impact on the results of model analysis. Considering the above rules, we then obtain a total of 3622 keywords for topic mining.

3.2 LDA Topic Analysis

The objective of LDA topic analysis is to examine the abstracts extracted from the papers. After performing word segmentation and part-of-speech tagging of the abstracts, we apply different word selection strategies to the results. We also set the prior distribution hyperparameters α and β to 0.1 and 0.1 respectively, and the number of topics is also set within the range of 5 to 60 with a step size of 5.

4 Result Analysis

The results of different word selection methods of the LDA topic analysis on the abstracts are presented in Tables 1 to 3. In these experiments, the number of topics is set to 10, 7 high-intensity topics are selected, and several high-probability words are picked.

Table 1. Results of “noun + verb” words selection

Topic No.	High Probability Words
1	research, education, student, development, learning, teacher, learner analysis, technology, teaching, application, design
2	development, culture, education, research, information technology, students, change, application, technology
3	service, construction, application, improvement, development, optimization, situation, promotion, interaction, informatization,
4	information technology, learner, interaction, precision, teacher, learning, problem, network, content, intelligence, ability
5	Children, school, intervention, cognition, performance, understanding, thinking, assessment, support, factor, form, skill, effect
6	framework, theory, wisdom, education, construction, informatization, application, practice, artificial intelligence, development, goal, curriculum
7	university, major, presentation, system, content, students, international, teaching, learning, bringing, organizing, network, form, space

Table 2. Results of "noun" words selection

Topic No.	High Probability Words
1	cyberlearning space, information technology, space, education information, management, organization, development, function, promotion, improvement, environment
2	steam education, program, reform, goal, computational thinking, guidance, quality, science, improvement, mechanism
3	smart classroom, wisdom, support, activity, utilization, background, optimization, informatization, network, group, policy, service
4	maker education, wisdom education, precision teaching, connotation, big data, culture, implementation, transformation, theory, reform, precision, and practice
5	teaching, subject, student, skill, knowledge, application, ability, perspective, research, discipline, foundation, thinking, education, learning
6	computational thinking, deep learning, MOOC, research, tools, theme, embodiment, core, stage, reflection, system, method, result
7	research, student, development, teacher, learner, educational analysis, problem, learning, technology, impact, application, artificial intelligence

Table 3. Results of key-words selection.

Topic No.	High Probability Words
1	online learning, interactive, cyber learning space, personalized learning, learning behavior, learning analysis, learning resources
2	artificial intelligence, big data, classroom teaching, education big data, artificial intelligence technology, wisdom, smart classroom, education and teaching, intelligent education, deep integration
3	deep learning, blended learning, experience, interaction, knowledge construction, generation, learning experience, information literacy, questionnaire
4	learning effect, elementary education, precision teaching, emotion, assessment, instructional design, educational games, data-driven, data analysis
5	computational thinking, programming, interdisciplinary, digital education resources, teaching effectiveness, measurement, core literacy, teaching methods, learning activities
6	steam education, smart education, learning environment, reading, virtual reality, educational games, new technology, statistics, education field, research, learning habits
7	maker education, educational technology, children, lifelong learning, Internet plus, elements, integration, diversity, interviews

As it is seen in Table 1 to Table 3 the keyword-based word selection scheme provides more specific and easy-to-interpret professional terms, making it easier for users to understand specific topics and high-probability vocabulary covered under the topics. To analyze the characteristics of the three-word selection strategies, we further use a combination of qualitative and quantitative methods to evaluate the results.

4.1 Qualitative Assessment

To evaluate the effectiveness of different word selection strategies, we adopt a qualitative evaluation method for preliminary analysis. We invite 7 professionals in the field of education technology to rate the topic terms (scores between 0 and 10). The scoring takes the form of an online questionnaire. Each question stem of the questionnaire is a topic word selected from the LDA model results. Each invited expert scores 20 sets of results. The scores represent the degree of conformity with the subjective cognition of experts in the research topics and related words in the sub-field of educational technology under each topic. The results of this experiment are presented in Table 4.

Table 4. Experts' rating of the results

No. Methods	1	2	3	4	5	6	7	Avg.
Noun+verb	2	1.4	1.2	1.7	1.6	1.3	1.8	1.6
Noun	3	3.1	2.4	3.3	2.7	2.4	3.4	2.9
Keywords	7.1	6.3	6.5	6.8	5.6	7.4	6.4	6.6

As it is seen in Table 4 the average score obtained by the keyword-based LDA topic analysis is higher than that of the other two-word selection strategies. Further interviews with the experts also confirm that the topic words generated by the first two-word selection methods include many common words and the readability and interpretability of the topics are generally poor. The keyword-based selection strategy however presents the topic words with strong readability, can summarize the core content of the topic, and the overall performance is better than the first two-word selection strategies.

We also acknowledge that having only 7 domain experts is probably not representative enough. It is also seen that in the keyword selection method, the scores under different topics are quite different. Some topics only got 4 points, while some other topics got 9 points. In the word selection method based on “nouns”, some topics also got scores from 5 to 6 points. This indicates the strong subjectivity of this qualitative analysis. The scoring analysis can only reflect the difference of the three-word selection methods to a certain extent and cannot provide a scientific and reasonable explanation of the analysis results. Therefore, we also perform quantitative analysis to further analyze the performance of the three types of word selection strategies and different topic models.

4.2 Quantitative Analysis of Word Selection Methods

The quantitative analysis methods of topic models can be generally divided into two categories including external (related to the task) and internal (not related to the task) methods. The extrinsic methods usually evaluate the quality of the model based on the performance of subsequent natural language processing tasks (such as the accuracy and recall rate of text classification, etc.). The intrinsic methods focus on evaluating the quality of the generated topic and are independent of the subsequent applications.

Here we analyze the performance of the LDA topic model based on the keyword selection strategies and do not consider the subsequent application of the model. Therefore, four internal methods of model perplexity, topic consistency, inter-topic distance are used in our analysis.

4.2.1 Model Perplexity

The perplexity of the model is used to measure the generalization ability of the model training results. Perplexity is the degree of certainty of the probability of topic classification in the test set. Let N denote the number of test documents, and N_d represents the number of words in the document, d. The model perplexity is then defined as

$$\text{Perplexity} = -\frac{1}{N}\sum_{d=1}^{N}\frac{1}{N_d}\log p(w) \qquad \textbf{(1)}$$

where p (w) denotes the probability of each word in the test document,

$$p(w) = p(z \mid d) * p(w \mid z) \qquad \textbf{(2)}$$

And p(z|d) is the probability of each topic in a document, d, and p(w|z) represents the probability of each word under a certain topic. These two probabilities are calculated by training θ and φ.

Fig. 2 shows the model perplexity results of the three-word selection strategies for a different number of topics. It is seen that the perplexity of the “keyword” selection method is the lowest amongst the three methods. The LDA model constructed based on the keyword selection method has a higher generalization performance.

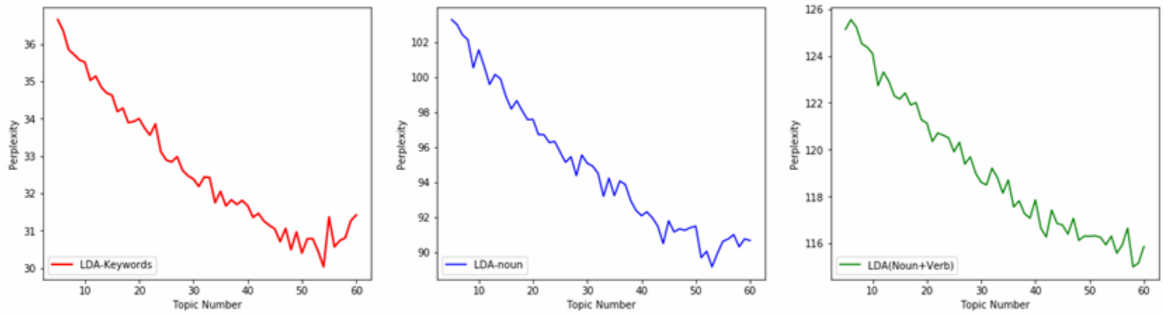


Fig. 2. The perplexity of three-word selection strategies with different numbers of topics

4.2.2 Topic Coherence Measure

Topic Coherence Measure refers to the calculation of semantic similarity scores between high-probability words within a topic. This score can be used to distinguish between better and less interpretable topics. Topic terms with better interpretability have a higher consistency score (the result is close to 0 from negative). The topic coherence calculation method is

$$\text{coherence}\,(T) = \Sigma_{(V_i,V_j)} score(V_i, V_j) \tag{3}$$

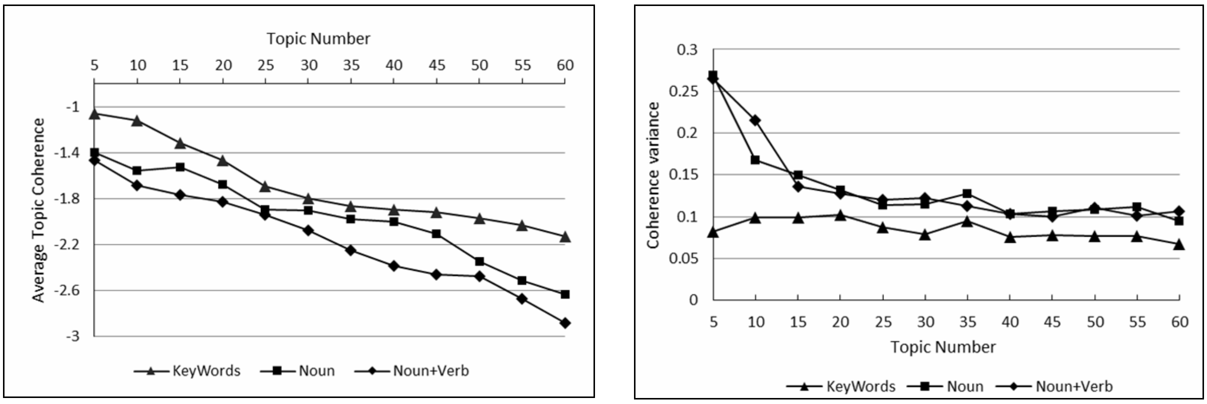
where (V_i, V_j) is a word pair under the topic. The value of the score function is obtained using the Umass algorithm, which calculates the word pair in the document. The co-occurrence probability measures its similarity

$$score(V_i, V_j) = \log \frac{D(V_i, V_j) + \in}{D(V_j)} \tag{4}$$

where $D(V_i, V_j)$ represents the number of documents containing the word pair, $D(V_j)$ represents the number of documents containing the word V_j, and $\in$ is a smoothing factor that ensures that the result of the operation is a real number. To reduce the impact of the smoothing factor on the scoring results, Mimno [19] sets $\in = 10^{-12}$.

It can be seen in Figure 3(a) that the average topic coherence values of the three-word selection methods show a downward trend by increasing the number of topics. This is because for a limited number of corpora, by increasing the number of topics, the number of documents that can be divided into the same topic is decreased. This decreases the co-occurrence probability of topic words. For the three different word selection methods, the average topic coherence of the keyword-based topic analysis method is greater than that of the “noun” and “noun + verb” methods.

From the variance statistics shown in Figure 3(b), for different numbers of topics, the variation of the topic coherence variance obtained based on the keyword method is relatively more stable and smaller than that of the other two methods.



(a) Average topic coherence

(b) Coherence variance

Fig. 3. Topic consistency analysis

4.2.3 Inter-topic Distance

The topic coherence measures the semantic consistency within the topic. Different topics need to maintain a certain distance to better divide the documents into topics. Therefore, we also measure the distances between topics (inter-topic distance). In topic model analysis, it is usually we expect a relatively large distance between the topics. This means a higher degree of distinction between the topics. If the distance between topics is too small (i.e., the topic vector similarity is too high), the topic analysis is overfitted. To measure inter-topic distance, here we use cosine distance. For the two topics T_a and T_b, the cosine distance is defined as

$$D_c(T_a, T_b) = 1 - S_c(T_a, T_b) \tag{5}$$

where D_c represents the cosine distance of the two vectors, and S_c represents the cosine similarity of the two vectors and

$$S_c(T_a, T_b) = \frac{\boldsymbol{T_a} \cdot \boldsymbol{T_b}}{\|\boldsymbol{T_a}\| \cdot \|\boldsymbol{T_b}\|} = \frac{\sum_{i=1}^{n} T_{ai} T_{bi}}{\sqrt{\sum_{i=1}^{n} T_{ai}^{\ 2}} \sqrt{\sum_{i=1}^{n} T_{bi}^{\ 2}}} \tag{6}$$

In (6) T_{ai} and T_{bi} are each component of the topic vector T_a and T_b respectively (represented by the normalized TF-IDF value of the topic word). Assuming that the number of topics is N, the cosine similarity between topics will form an N*N symmetric matrix (diagonal element value is 1, as shown in Fig. 4).

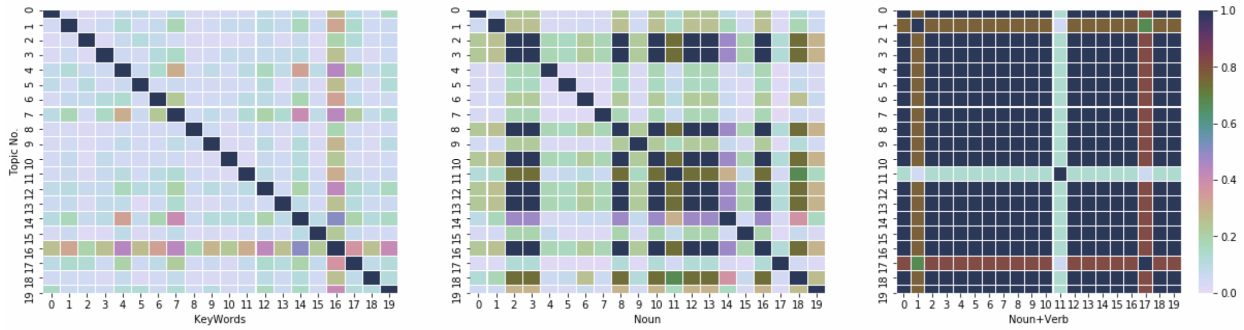


Fig. 4. Heat maps of cosine similarity matrix between the topics, where the number of topics is set to 20

The heat maps of the cosine similarity matrix between topics are constructed according to formula (6) when the number of topics is 20 (see Fig. 4). It is seen from Fig. 4 that the similarities between the three-word selection methods are different. The highest similarity between the topics is the word selection method based on "noun + verb" which represents overfitting. The least similarity is our proposed keyword selection method.

To comprehensively measure the similarity between the topics of the three word selection methods under different topic numbers, we calculate the average cosine distance between the topics for the topic numbers varying from 5 to 60 with a step size of 5.

As it is seen in Fig. 5, for different numbers of topics, better inter-topic distances are maintained between the topics obtained based on the keyword selection method. Also, by increasing the number of topics, the keyword-based method maintains a more stable cosine distance. Therefore, as it is confirmed by the above quantitative analysis, the topic analysis based on keyword selection achieves relatively high performance in terms of model perplexity, topic coherence, and inter-topic distance.

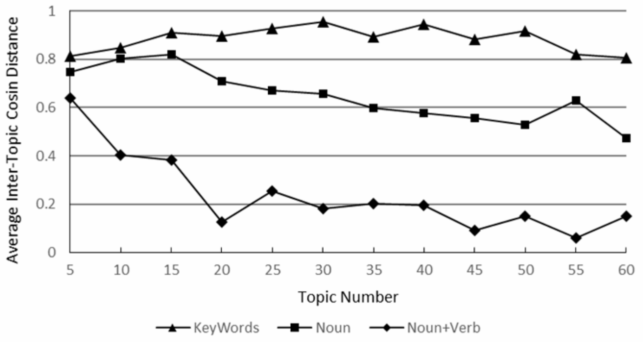


Fig. 5. Average cosine distance between the topics

4.3 Quantitative Analysis of Different Model Optimization Methods

4.3.1 Model Realization

For the selected three models (as shown in Fig. 1), we refer to the relevant implementation methods on the open-source community Github and make corresponding modifications. Since the source code of the three models processed and analyzed English corpus, the natural language processing part is modified so that the three models can process the data source selected in this paper. We also note that the GloVe pre-training word mapping matrix that is used in LDA-Word2Vec is not relevant here. Therefore, we use the gensim.word2vec module to pre-train the data with the continuous bag-of-words model to generate the word vector mapping matrix and establish the semantic vector relationship between words.

We also set the same values for similar parameters in the compared models, for example, the hyperparameters α and β are fixed at 0.1, 0.1. The variation range of the number of topics is between 5 and 60. For the second level truncation of the HDP model, the value range of T is also 5-60. The change of the T value affects the optimal number of topics for HDP model analysis. Based on this, the model simulates the changes in the number of topics.

4.3.2 Model Perplexity

Fig. 6 illustrates the model perplexity of different model optimization methods. It can be seen from the figure that the LDA topic analysis method based on keyword selection (LDA-Keywords) has the lowest degree of perplexity, followed by the LDA-Word2Vec method. After the corpus is screened by the keywords in the abstracts, the scope of corpus words is narrowed down, and high-frequency common words or low-frequency rare words in the domain are removed. The results of the analysis can be more focused on the topics constituted by the keyword set, hence obtaining a lower degree of perplexity than that of the other models.

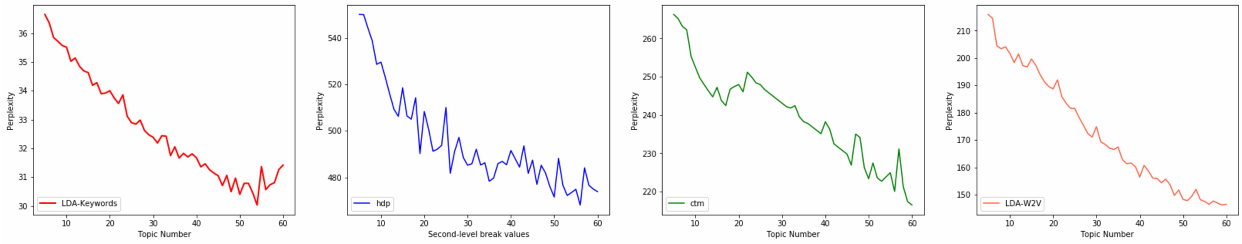


Fig. 6. Model perplexity for different topic models

Note that amongst the four compared methods, the LDA analysis based on the word-vector model improves the semantic relationship between words. With the help of the word embedding matrix, words with close semantic distance are distributed into the same topic with high probability. As a result, the model perplexity is significantly better than that of the HDP and CTM.

For the HDP model, the main feature is dynamically determining the number of topics. Its estimation of document-topic and topic-word probability distribution however is based on the Dirichlet model. Therefore, as it is also seen in Fig. 6 when selecting different second-level break T values to simulate the change of the number of topics, the model perplexity value is higher than that of the other three models.

4.3.3 Predictive Log-likelihood

In addition to the direct comparison of model perplexity, several models selected in the paper are compared according to the model measurement method adopted in [20]. For the document data to be analyzed, the model selects 10% (about 1300 documents in this article) of the words from the test set D_{test} . The remaining 90% of the documents are used as the training set D_{train} for the model parameters. We also calculate the predictive log-likelihood of the document ω_j in the test set D_{test} under the condition that D_{train} is a priori. This avoids direct comparison of hyperparameters of different models:

$$log - likehood_{pw} = \frac{\Sigma_{j \in D_{test}} \log p(\omega_j \mid D_{train})}{\Sigma_{j \in D_{test}} \mid \omega_j \mid} \quad \textbf{(7)}$$

where $\mid \omega_j \mid$ represents the number of different words in ω_j. It is more difficult to directly calculate the conditional probability in (7). Therefore, Wang et al. [20] suggested using the approximate calculation method as

$$p(\omega_j \mid D_{train}) \approx \prod_{\omega \in \omega_j} \Sigma_k \overline{\theta}_{jk} \overline{\phi}_{k\omega} \quad \textbf{(8)}$$

In (8) k represents the number of topics. Intuitively, the predicted log-likelihood value of the document set ω_j is obtained by multiplying the predicted likelihood value of its word. The predicted likelihood value of each word is expressed by the sum of the probability distribution values of the words. The word probability score is then calculated by multiplying the average probabilities of the document-topic (θ) and the average topic-term (ϕ) probability of the document where the words are located.

This measurement is preferred because it does not need to directly compare the hyperparameters of different models. Instead, it comprehensively measures the performance of the model hyperparameters by calculating the likelihood value of the test set.

If the effect of model hyperparameter training is higher, the above formula (7) returns a larger value with a high probability. Correspondingly, if the model parameters are overfitted, this inevitably leads to a large number of small probability distributions in the document-topic and topic-word matrix. Therefore, on the test set, there is a high probability of a small likelihood value (the values are all negative).

After training the models using D_{train} document sets, four different methods' prediction log-likelihood values are calculated by using the D_{test} document sets (see Fig. 6). It is seen from the figure that when the number of topics is less than 40, the LDA-Keyword model's prediction log-likelihood value is better than that of the other three methods. Due to the impact of model parameter fitting, initially, the prediction log-likelihood values of all four models increase with the increase of the number of topics. Beyond a certain number of topics, however, the HDP and CTM models show an over-fitting phenomenon hence the likelihood value shows a downward trend. The LDA-W2V is slightly better than that of our proposed method, however, training the word2vec model is more complex than that of the keywords selection.

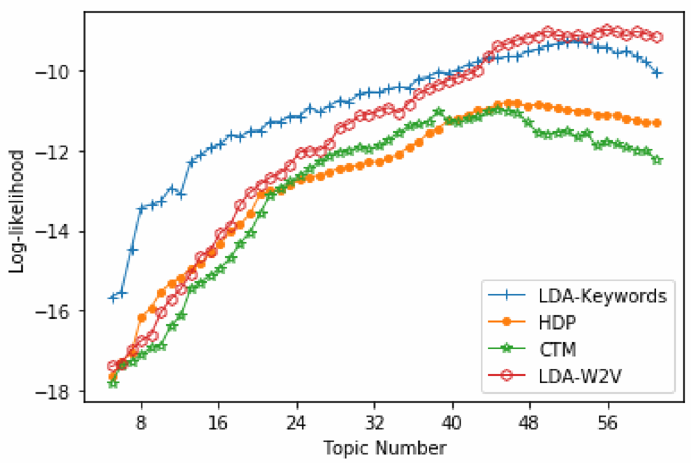


Fig. 7. The predictive log-likelihood value versus the topic number

5 Conclusion

Aiming at the problems of weak model generalization ability and poor interpretability of topic words of the LDA topic model, in this paper we designed a topic analysis framework based on keywords selection. We then performed qualitative and quantitative performance evaluations and compared them with different word selection strategies and different topic models described in the paper, the proposed method was shown to be easier to implement and have a higher performance. In the LDA topic analysis, the results of general models are often quite different when they are applied in different fields. Therefore, in future research, we further incorporate and analyze the characteristics of different application domains and design a topic model that fits better to the domain characteristics.

Acknowledgements

This research was funded by the joint project from the Henan-Provincial and the China-National Natural Science Foundation's under grant U1804153, the project of science and Technology Department of Henan Province (192102310288). The authors would like to express their gratitude to EditSprings (https://www.editsprings.com/) for the expert linguistic services provided.

References

[1] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, Journal of Machine Learning Research 3(2)(2003) 993-1022.

[2] B.-G. Yu, Y.-M. Cao, Y.-N. Chen, Classification of Short Texts Based on nLD-SVM-RF Model, Data Analysis and Knowledge Discovery 4(1)(2020) 111-120.

[3] J.-J. Huang, P.-W. Li, M. Peng, Review of Deep Learning-based Topic Model, Chinese Journal of Computers 43(05)(2020) 827-855.

[4] K. Stevens, P. Kegelmeyer, D. Andrzejewski, D. Buttler, Exploring Topic Coherence over Many Models and Many Topics, in: Proc. 2012 Proceedings of empirical methods in natural language processing, 2012.

[5] J.D. Lafferty, D. M. Blei, Correlated topic models, in: Proc. 2006 Proceedings of the International Conference on Neural Information, 2006.

[6] C. Yau, A.L. Porter, N.C. Newman, Clustering scientific documents with topic modeling, Scientometrics 100(3)(2014) 767-786.

[7] J. Yin, J. Wang, A Dirichlet multinomial mixture model-based approach for short text clustering, in: Proc. 2014 Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2014.

[8] R. Muzafar, A.K. Majid, A.T. Tanveer, Deep LDA: A new way to topic model, Journal of Information and Optimization Sciences 5(2)(2019) 101-112.

[9] M. Peng, S.-X. Yang, J.-H. Zhu, Semantic Enhanced Topic Modeling by Bi-directional LSTM, Journal of Chinese Information Processing 32(4)(2018) 40-49.

[10] T.-T. Wang, M. Han, Y. Wang, Optimizing LDA Model with Various Topic Numbers: Case Study of Scientific Literature, Data Analysis and Knowledge Discovery 2(1)(2018) 29-40.

[11] R. Das, M. Zaheer, C. Dyer, Gaussian LDA for Topic Models with Word Embeddings, in: Proc. 2018 International joint conference on natural language processing, 2018.

[12] R. Ding, R. Nallapati, B. Xiang, Coherence-aware neural topic modeling, in: Proc. 2019 Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2019.

[13] G. Chen, W. Wu, Comparative Study on the Effect of Word Selection Scheme on the Effect of LDA Analysis in Subdivisions, Information Studies:Theory & Application 42(6)(2019) 25-31.

[14] H. Lei, Y. Chen, Exclusive Topic Modeling, <http://arxiv.org/abs/2102.03525>, 2021(accessed 20.02.21).

[15] Y. Zhang, G. Zhang, H. Chen, Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research, Technological Forecasting and Social Change 105(2)(2017) 179-191.

[16] Y. Zhang, H. Chen, J. Lu, Detecting and predicting the topic change of Knowledge-based Systems: A topic-based bibliometric analysis from 1991 to 2016, Knowledge Based Systems 133(2)(2018) 255-268.

[17] X.-C. Gong, X.-Y. An, A Research of Topic Splitting and Merging Detecting in the Medical Field Based on the LDA Model, Library and Information Service 61(18)(2019) 76-83.

[18] Q.-Y. Liu, Y. Ye, A Study on Mining Bibliographic Records by Designed Software SATI: Case Study on Library and Information Science, Journal of Information Resources Management 5(01)(2012) 50-58.

[19] P. Abels, Z. Ahmadi, S. Burkhardt, Focusing Knowledge-based Graph Argument Mining via Topic Modeling, <https://arxiv.org/abs/2102.02086>, 2021(accessed 20.03.21).

[20] C. Wang, J. Paisley, D.M. Blei, Online Variational Inference for the Hierarchical Dirichlet Process, in: Proc. 2011 International conference on artificial intelligence and statistics, 2011.