Research on the Evolution of Journal Topic Mining Based on the BERT-LDA Model

Guofeng Tang[1], *Xuhui* Chen [2,*] *Ning* Li [3] *and Jianfeng* Cui [4]

[1]College of Computer and Information Engineering, Xiamen University of Technology, Xiamen, 361024, China
[2]Fujian Key Laboratory of Internet of things application technology, Xiamen City University, Xiamen, 361008, China
[3]Editorial Department, Xiamen University of Technology, Xiamen, 361024, China
[4]College of Software Engineering, Xiamen University of Technology, Xiamen, 361024, China

Abstract. Scientific papers are an important form for researchers to summarize and display their research results. Information mining and analysis of scientific papers can help to form a comprehensive understanding of the subject. Aiming at the ignorance of contextual semantic information in current topic mining and the uncertainty of screening rules in association evolution research, this paper proposes a topic mining evolution model based on the BERT-LDA model. First, the model combines the contextual semantic information learned by the BERT model with the word vectors of the LDA model to mine deep semantic topics. Then construct topic filtering rules to eliminate invalid associations between topics. Finally, the relationship between themes is analyzed through the theme evolution, and the complex relationship between the themes such as fusion, diffusion, emergence, and disappearance is displayed. The experimental results show that, compared with the traditional LDA model, the topic mining evolution model based on BERT-LDA can accurately mine topics with deep semantics and effectively analyze the development trend of scientific and technological paper topics.

1 Introduction

In the field of scientific research, papers are one of the important means for academic researchers to analyze and grasp this field [1]. However, with the development of the times, the types of journals and the number of papers are increasing. It is very important to utilize emerging technologies to mine and analyze hot topics in the scientific fields, to assist researchers in analyzing the hot topics in the industry, to provide a basis for researchers' research in a specific field, as well as to help scholars to find the direction of disciplinary research and let them understand the topics in this field. Governments and enterprises can also focus on the research hotspots in various industries, from which they can find the latest research direction and grasp the first opportunity.

Existing scientific paper research mainly uses technologies such as keyword analysis [2],

* Corresponding author email: xhchen@xmut.edu.cn

citation analysis [3], and topic analysis [4] to mine paper topics, while using topic evolution to analyze changes in topic hotspots under time windows and to demonstrate process phenomena such as birth, growth, splitting, convergence, decline, and extinction among themes. However, existing studies do not remove the interference of edge texts during data pre-processing, as well as topic mining with less analysis of the semantics of the text's context, treating the text as a collection of words only. When filtering the topic association, it is often based on the author's experience, and the subject association filtering rules are rarely used.

Aiming at the shortcomings of the present research, this paper first uses a sentence transformer [5] to generate the hottest text set. Then BERT is used to learn the contextual semantic information of the text, which is combined with LDA to mine the topic of deep semantic information. Finally, the topic association filtering rule is established to remove unimportant topic associations and construct the evolution path to analyze the changing direction of industry research.

2 Relevant theoretical models

Topic mining is a technique that can mine hidden topic hotspots from large and complex texts, analyzing topic changes in the field. With the development of technology in the field of topic mining, the existing topic mining methods mainly include co-word analysis, word frequency analysis, text clustering, topic modeling, algorithm improvement, etc. Chen et al. [6] used Python to count the frequency of keywords in journals related to text sentiment in CNKI and used SPASS software to do the co-word analysis of keywords. Combined with multiple scales analyze the current situation of text emotion in China in recent years and reveal the research hotspots in this field. Shang [7] used word frequency statistics to mine the themes of the National intelligence law of the people's Republic of China, combined with social network analysis software to connect the relationship between themes and various articles, and analyzed the operating mechanism behind the national intelligence agencies. Feng et al. [8] introduced a time-weighted approach based on the word frequency statistics method to assign time weights to the mined keywords, which can dynamically track and analyze the subject hotspots of the subject. The methods of vocabulary-based statistics mostly use word frequency statistics or co-word analysis, which have problems such as word frequency threshold and mining the association between words. They are only applicable to small-scale data analysis, and the method is relatively simple [9]. Wang et al. [10] used Kmeans to cluster privacy log data to analyze user privacy behavior and formulate rules to constrain user violations. Text clustering is highly subjective in labeling categories and is influenced by noise, which is not suitable for complex text analysis. Li et al. [11] used the LDA model to mine the topic of the Chinese Financial Stability Report to analyze the financial trend. Wang et al. [12] used LDA to mine the like topics of Weibo users as a predictor of user characteristics. Topic model LDA can be combined with the life cycle, time series, and other methods to analyze the changing process of topics in various periods in the research field, which is suitable for large text analysis, but they ignore the semantic information of the context and cannot express the gist of the context of the article. Zhang et al. [13] used a convolutional neural network to extract the deep-level features of news texts based on word2vec, combined with K-means clustering to mine news topics. By comparing with the single clustering method, they found that the improved precision and recall rate of this algorithm is higher than that of the single clustering method. Ruan et al. [14] used the Doc2Vec algorithm to learn the contextual semantics of abstracts and combined it with kmeans++ clustering to mine popular topics in journal abstracts. Hu et al. [15] combined LDA and Doc2Vec algorithms to mine policy text topics containing contextual semantic information to analyze and interpret policies, which helps people

understand policies. Although Doc2Vec considers the order of words and can predict the missing words in the sentence during pre-training, it does not solve the problem of polysemy. In contrast, Bert uses the transformer feature extractor to extract words and learn contextual semantic information to solve the problem of polysemy [16].

The topic evolution is the process by which themes change over time, demonstrating the path and direction of this discipline, and plays an important role in discipline research [17]. Topic evolution can display research hotspots and research status in the field through visualization tools, which is convenient for researchers to analyze and understand useful information [18]. At present, there are many visualization software for topic evolution. In 2006, Chen [19] proposed the visualization software CiteSpace, which can display the trend of topics changing over time. In 2011, Caboz [20] proposed the visualization software Sci MAT, which can show complex relationships between subject terms. In 2013, Wang [21] proposed the Neviewer visualization software, which can display the complex evolution of the topic structure. Wu et al. [22] used SciMAT to draw the relationship between topics in the online medical field over time. Niu et al. [23] used co-word analysis to construct topic networks in different periods according to journal keywords and used Gephi software to construct topic evolution paths. Although the topic evolution visualization software can display evolution well, evolution is a complex process, and it cannot analyze the evolution process from a multi-dimensional perspective. Li [24] used the LDA model combined with the post-discrete-time approach to mine the epidemic topics in the WeChat public account, and constructed and analyzed evolutionary paths in terms of three dimensions: topic intensity, topic attention, and topic identity. Zhu et al. [25] used LDA to mine topics and designed three dimensions of topic novelty, topic hotness, and topic migration as predictors. Jensen-Shannon between adjacent topics is calculated to construct topic evolution. However, when filtering topic associations, the authors set thresholds to screen the correlation between topics according to their own experience, without fixed screening rules. Yan et al. [26] firstly used the LDA model to mine scientific papers in the graphene field and then established topic association rules to screen topic associations in adjacent periods. Finally, they constructed evolution paths to analyze frontier science and technology in the graphene field. However, the screening rules still have the author's own experience and are not fully systematic, explaining all the screening thresholds clearly.

To sum up, the existing topic of mining evolution models still has shortcomings. First, the interference of edge text is not removed in the experiment. Second, the contextual semantic information of the text and the semantic order of words are ignored in topic mining. Third, most of the existing topic association screening is based on the author's own experience, and there is no set of mature topic relevance screening rules.

Based on the above research, the main contributions of the BERT-LDA topic mining evolutionary model proposed in this paper are as follows.

(1) This paper uses the sentence-transformer to generate the hottest text set to remove the interference of edge text on topic mining.

(2) The BERT model is used to learn the contextual semantic information of scientific papers, and combine it with the LDA model to mine deep semantic information.

(3) In the topic evolution layer, this paper further improves on the previous research by constructing topic association screening rules to filter the association relationship between topics and eliminate invalid topic associations.

3 Bert-LDA topic mining evolutionary model

This paper proposes a BERT-LDA topic mining evolutionary model, which includes three layers: data pre-processing layer, topic mining layer, and topic evolution layer. The data pre-processing layer consists of five steps: text de-duplication, jieba splitting, lexical

filtering, removing stop words, and using a sentence transformer to remove edge text. In the topic mining layer, the BERT model is combined with the LDA probabilistic topic mining model to learn contextual semantic information, and the Encoder and Decoder are introduced to mine deep topics. In the topic evolution layer, the topic association filtering rules are set to eliminate the invalid themes in the themes of each year, and analyze the topic evolution pattern from two dimensions topic hotness and topic migration degree. This paper proposes a research framework as shown in Figure 1.

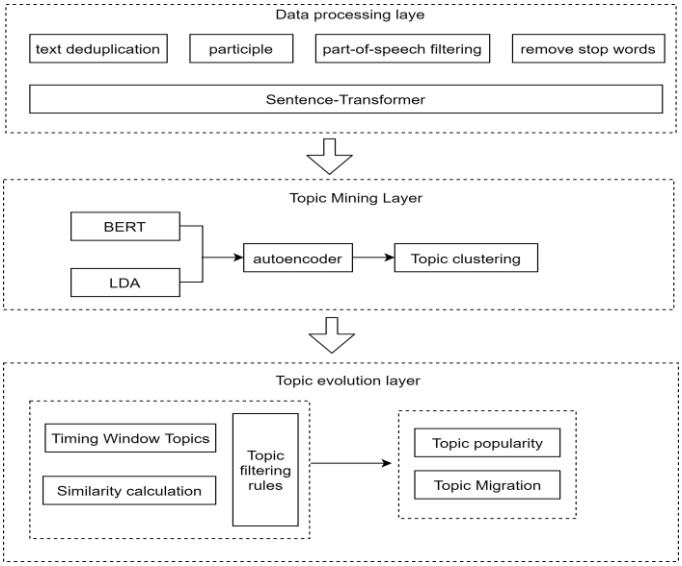

Fig. 1. BERT-LDA topic mining evolutionary modeling framework

3.1 Data pre-processing layer

Data pre-processing is the basis of the entire BERT-LDA topic mining evolution model. Preprocessing the original data obtained by the crawler is an important means to reduce the interference of incorrect texts on topic mining and to improve the accuracy of topic mining. The data pre-processing of this model mainly includes the following steps: 1) Use sentence-transformer to remove the interference of edge journals; 2) Text deduplication; 3) Jieba splitting, lexical filtering, and stop word removal, etc.

3.2 Topic mining layer

The LDA topic model is based on the "document-topic-word" 3-layer Bayesian model. Each document will generate multiple topic probabilities that sum to one, and each topic is composed of the probabilities of multiple words that sum to one [27-28]. Since the LDA model ignores the contextual structure of text when mining textual topics, this paper introduces the BERT model to learn the contextual semantic information of the article to make up for the insufficiency of the bag-of-words model.

BERT combines the idea of bi-directionality in ELMo and the use of transformers in GPT. BERT does not need to use the annotated corpus and uses the mask to replace part of the original data training corpus to learn contextual semantic information. The sentence-level representation of BERT is shown in Figure 2. The input of BERT consists of three parts: Token Embeddings, Segment Embeddings, and Positional Embeddings. Different

Tokens have different semantics. Segment Embeddings are used to distinguish two sentences, and different sentences have different Segment Embeddings. Since the entire sentence is input into BERT training and the position of each word is unknown, Positional Embeddings are added to confirm the position of each word.

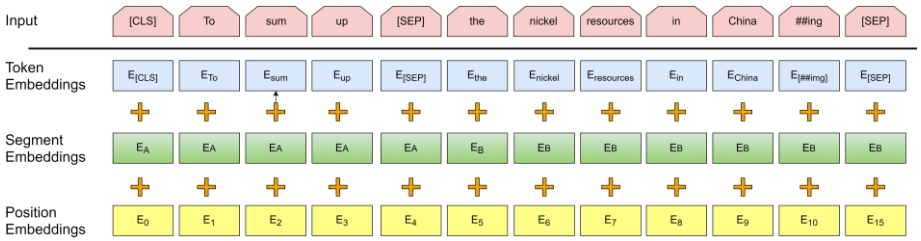

Fig. 2. Sentence-level representation of BERT

The BERT-LDA model is shown in Figure 3. Step1: Input the text into the LDA model and the BERT model respectively to generate the topic probability vector and the sentence vector containing the contextual semantic information. These two vectors are fused and input to the autoencoder; Step 2: In this paper, the autoencoder is used to perform feature dimensionality reduction, learn the low-dimensional spatial representation of the connected vector, and cluster the learned low-dimensional parameters.

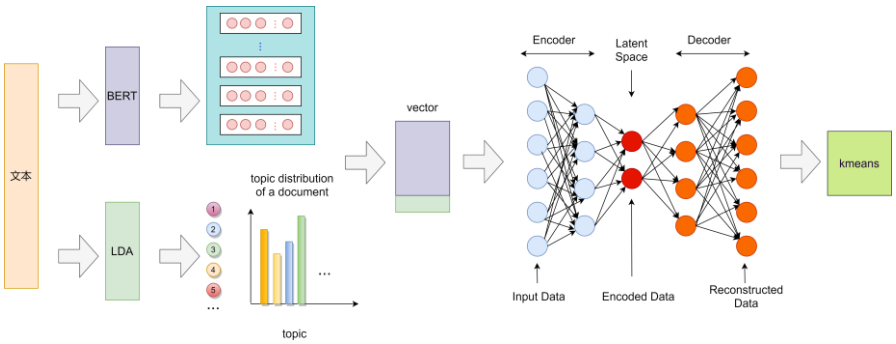

Fig. 3. BERT-LDA model

3.3 Topic evolution layer

3.3.1 Topic association

There are continuations and changes of topics between papers published in scientific journals every year. This paper shows the continuing relationship between themes and establishes thematic associations by calculating the similarity between the themes of two adjacent windows.

Firstly, in this paper, all the topics at moment $t+1$ are similarly calculated to the topic T_i at moment t. The backward topic with the maximum similarity at moment $t+1$ is taken as $t(T_i)$.

$$t+1(T_j) = post(t(T_i)) \quad (1)$$

The similarity of all topics at moment t to the topic T_i at moment t+1 is calculated, and the forward topic whose maximum similarity at time t is $t+1(T_j)$ is taken.

$$t(T_i) = \text{prior}(t+1(T_j)) \quad (2)$$

3.3.2 Topic filtering rules

Because when doing topic association evolutionary associations, some topics have a low similarity or even no similarity, a topic filtering rule is needed to eliminate subjects with low subject relevance. In this paper, three topic association filtering rules are used. If any one of them is satisfied, the association is regarded as invalid.

The topic association filtering rules in this article are as follows:

(1) Take the topic in the t period as T_i, and all topics in the $t+1$ period as T. Calculate the average of the similarity between the topic T_i in the t period and all topics T in the $t+1$ period. If the topicT_i and topic T_j the topic similarity is less than the average value, the association is judged invalid, and vice versa.

(2) $t+1(T_j)$ is the backward topic of $t(T_i)$. Calculate the similarity between all topics at time t and $t+1(T_j)$ to take the topic with the highest similarity at time t as T_k. Take the topic T_n whose similarity is between T_k and $t(T_i)(i>2)$. If the backward topic of T_n is not $t+1(T_j)$, the topic evolution relationship between them is invalid.

(3) $t(T_i)$ is the forward topic of $t+1(T_j)$. Calculate the similarity between $t(T_i)$ and all topics at time $t+1$, and take the topic at time $t+1$ with the highest similarity as T_k. Take the topic T_n with the similarity between T_k and $t(T_i)(i>2)$. If the forward topic of T_n is not $t(T_j)$, the topic evolution relationship between them is invalid.

3.3.3 Dimensions of evolutionary analysis

(1) Theme popularity

Topic popularity reflects the popularity of the topic in the current period, and the popularity of the topic is represented by the proportion of the number of journal papers under the topic to the number of journal papers used in this period.

$H_t(T)$ represents the popularity of topic T at time t; $\text{Sum}(t(T_i)$ denotes the sum of the number of journal papers under the ith topic T at time t; $\text{Sum}(t(T_{all})$ denotes the sum of the number of journal papers under all topics at time t.

$$H_t(T) = \frac{\text{Sum}(t(T_i))}{\text{Sum}(t(T_{all}))} \quad (3)$$

(2) Topic Migration

Topic migration degree refers to the probability that a topic migrates from one window to the next, which encompasses the processes of topic birth, growth, division, fusion, decline, and demise. This paper uses text similarity to calculate probabilities between topics and uses topic filtering rules to filter relationships between topics.

4 Experiment and result analysis

4.1 Dataset acquisition

The data set in this paper comes from 30 core journals related to geological research on

CNKI. The selenium framework is used to crawl the journal name, paper title, abstract, publication time, and keywords of a total of 43,192 scientific papers from 2010 to 2020.

4.2 Text preprocessing

In this paper, basic operations such as jieba word segmentation, removal of modifiers, and removal of stop words are performed on the crawled data. Since the geological field data contains a large number of proper nouns, to prevent the proper nouns from being segmented during jieba word segmentation, this paper uses the keyword as a proper noun dictionary.

Since the topics of some papers in the data may only be short-lived and cannot be regarded as hot spots of research at that time, this paper uses sentence-transformer [29] to remove edge text and improve the accuracy of topic mining. When inputting a certain paper data, if similar paper data set is found scattered in multiple journals, it can be regarded as the hottest text. In the experiment, the top 100 papers with the most similarities were selected as the collection of papers. If the collection is scattered in 17 journals, the paper is kept in the hottest text set. The number of the hottest text sets in each time window is as follows shown in Table 1.

Table 1. The number of hottest texts

time	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
original quantity	4076	3989	4363	4292	4165	4067	3625	3701	3380	3754	3780
Later quantity	3879	3879	4128	4189	3996	3435	3384	3407	3129	3272	3477

4.3 Topic mining

4.3.1 Determine the number of topics

The number of topics is a key parameter of the topic model. When the number of selected topics is different, the clustering quality of the topic model is also different. There are two main indicators for evaluating the quality of the existing model: perplexity and consistency. However, since the topic perplexity will overfit when the number of topics is large, this paper uses Coherence (topic consistency) to determine the optimal number of topics for the BERT-LDA model. The greater the topic consistency, the better the model fit. After many experiments in this paper, the optimal number of topics is selected, as shown in Figure 4.

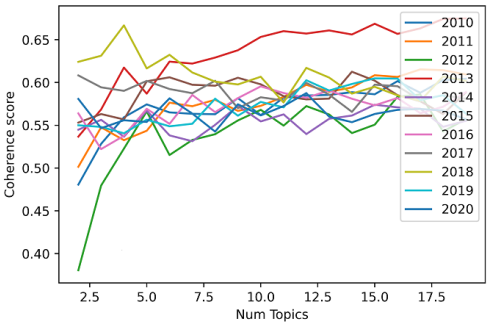

Fig. 4. Number of best themes by year

In this paper, topic consistency is used to select the best number of topics in different

periods, and the topics are screened to eliminate topics that are not related to the geology of this paper. For each topic, the top ten subject terms are selected for topic representation. Limited to the length of the paper, only the results generated by the theme in 2010 are shown in Table 2.

Table 2. Geological subject terms in 2010 based on the BERT-LDA model

time	Topic number	topic
2010	11	magma; element; geochemistry; rock; enrichment; mantle; granite; rock mass
		abnormal; geology; evaluate; prospecting; predic; resource; survey; work; china
		hydrate; simulation; natural gas; structure; parameter; a stratum; earthquake
		soil; groundwater; mineral; experiment; time; pollution; standard
		zircon; age; granite; magma; pb; ma; rock mass; structure
		deposition; environment; isotope; sediment; climate; biology; carbon; soil; profile
		deposition; basin; reservoir; sequence; oil and gas; structure; a stratum; sag; system
		earthquake; groundwater; rupture; fault; wenchuan; structure; crust; fracture; surface
		deposit; metallogenic; fluid; pack; porphyry; gold; parity; copper
		structure; fracture; basin; deformation; deposition; activity; night; fault; evolution
		loess; sediment; glacier; age; granularity; deposition; profile; relation

4.3.2 Comparative experiment

The BERT-LDA model proposed in this paper is compared with the traditional machine learning LDA model, and the Coherence (topic consistency) is used as an indicator of how good the model is. When the consistency of the new model is greater than that of the traditional model, it means that the new model has a better topic mining ability.

The period was divided according to year, and a certain period is randomly selected to compare the consistency of the BERT-LDA model and the traditional machine learning LDA model with the change in the number of topics. The results are shown in Figure 5. Under the same number of topics, the consistency of the BERT-LDA model is greater than that of the LDA model, indicating that the model proposed in this paper has better performance than the traditional model.

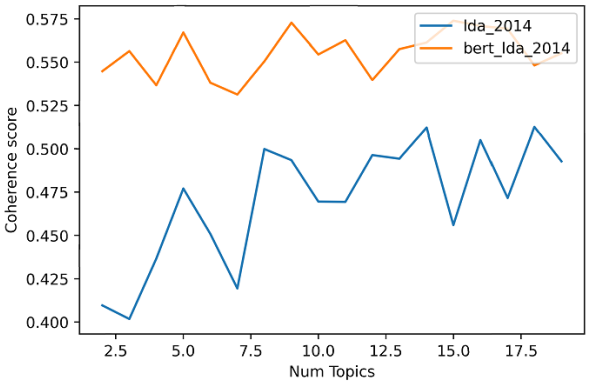

Fig. 5. Model comparison

In this paper, five articles are randomly selected to read and analyze their topics, and verify the accuracy of these five articles in mining themes in the BERT-LDA model and the LDA model respectively. The results are shown in Table 3 below.

Table 3. This is a table. Comparison of the BERT-LDA model and LDA model

id	title	LDA model	BERT-LDA model	accuracy
1	Temperature distribution and controlling factors of the Tangquan geothermal field in Hebei Province	groundwater; sediment; increase; landslide; concentration; trend; debris flow; temperature; relation	soil; groundwater; provide profiles; my country; system; temperature; sediment; the difference	almost
2	Prospecting thinking and model for the Ketinghaer porphyry copper-molybdenum deposit in the East Kunlun Mountains	abnormal; work; china; geology; provide; evaluate; propose; my country; resource; Predict; divide; achievement	deposit; ore body; metallogenic; ore-forming fluid; pyrite; quartz; mining area; ore; skarn	BERT-LDA
3	An analysis of the activity of the northwest part of the Nankou-Sunhe fault	basin; fracture; structure; fault; source rock; oil and gas; profile; activity; deposition; direction; lie in	fracture; earthquake; structure; age; activity; fault; profile; deformation; fracture; performance	BERT-LDA
4	Geochemistry and zircon U-Pb geochronology of the diorite associated with the Wuling iron deposit in Western Tianshan Mountains, Xinjiang	granite; rock mass; enrichment; loss; zircon; rare earth elements; element; diorite; rock	zircon; age; granite; rock mass; enrichment; loss; low; diorite; element	BERT-LDA
5	Lithofacies-paleogeography of middle-late Ordovician Daping stage-Aijiashan stage on the western margin of the Ordos Basin	deposition; reservoir; sandstone; crack; limestone; type; stratum; mudstone; porosity; dolomite	deposition; inclusions; salinity; ore-forming fluid; NaCl; sequence; profile; quartz; co2; basin; delta	BERT-LDA

4.4 Evolution analysis

4.4.1 Topic heat analysis

The three main themes are selected to plot the rate of change in topic analysis, as shown in Figure 6. It can be seen from Figure 6 that the topic's popularity changes significantly over time. The changes in the environmental pollution theme among these three themes are relatively gentle, which shows that my country has been conducting in-depth research on environmental pollution-related aspects. In 2016, the State Council issued the "Thirteenth Five-Year Plan" to promote sustainable and healthy economic development. People pay more attention to environmental research issues, and environmental pollution has also increased. The subject of geological movement (earthquakes, landslides, and debris flows) is also a subject that geological researchers have been studying. The popularity of sedimentary geology topics has been high, basically maintained above 0.1, with the popularity peaking in 2015.

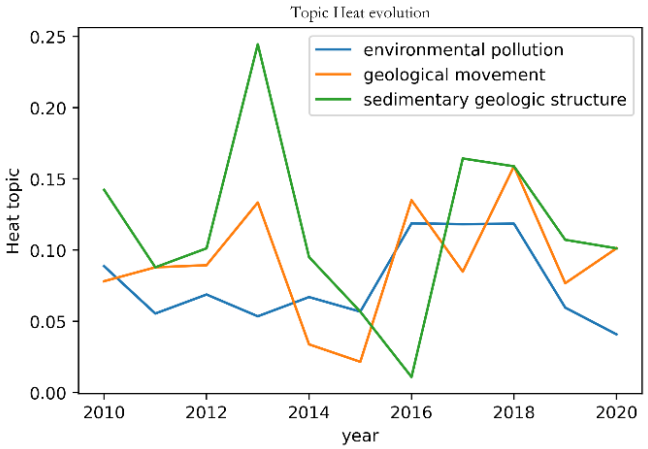

Fig. 6. Changes in topic popularity

4.4.2 Topic migration analysis

To study the relationship between topics, BERT-LDA is firstly used to mine topics in each period, then text similarity is used to calculate the similarity of adjacent Windows, and finally, topic filtering rules are used to filter and remove the topic connections that do not meet the conditions. Use the themes from 2013 and 2014 as a demonstration to show the thematic connections between them.

(1) Topic Filtering Rules 1

This paper calculates the average of the similarity of all topics from 2013 to 2014 as the threshold for transfer between topics. The similarity between topic 1 in 2013 and all topics in 2014 is shown in Table 4. T The average of all the similarities is calculated as the threshold value of 0.8361, and the associations whose similarity exceeds the threshold are retained, and vice versa.

Table 4. Correlations between theme 1 in 2013 and themes in 2014

2013	2014	similarity	average value	Whether the topic association is removed
Topic 1	Topic 1	0.8083	0.8361	no
	Topic 2	0.7029		no
	Topic 3	0.7860		no
	Topic 4	0.9261		yes
	Topic 5	08709		yes
	Topic 6	0.7576		no
	Topic 7	0.7767		no
	Topic 8	0.7793		no
	Topic 9	0.8752		yes
	Topic 10	09339		yes
	Topic 11	0.8377		yes
	Topic 12	0.7973		no
	Topic 13	0.8771		yes
	Topic 14	0.8571		yes
	Topic 15	0.8943		yes

(2) Topic Filtering Rules 2

Using rule 2 based on rule 1, backward themes were calculated for all topics in 2013, and the results are shown in Table 5. According to rule 2, the similarity between all topics in 2013 and backward topics is calculated, they are sorted according to the highest to the lowest. For example, the similarity between topic 8 and backward topic 9 in 2013 is 0.8191, the ranking position is 5th (5<x<1), and the backward topic of the middle 4 topics is not topic 9, then delete the middle 4 topics association with backward topic 9.

Table 5. Relationship between themes and backward themes in 2013

2013 theme	Post topic (2014 theme)	similarity	Sort Maximum Topics	similarity	Is the rule valid
Topic 1	Topic 10	0.9339	Topic 1	0.9339	no
Topic 2	Topic 1	0.9126	Topic 4	0.9216	no
Topic 3	Topic 1	0.8742	Topic 4	0.9216	no
Topic 4	Topic 11	0.9473	Topic 4	0.9473	no
Topic 5	Topic 6	0.8945	Topic 5	0.8945	no
Topic 6	Topic 13	0.9186	Topic 6	0.9186	no
Topic 7	Topic 1	0.8937	Topic 4	0.9216	no
Topic 8	Topic 9	0.8191	Topic 6	0.8985	Topic 8 <x< Topic 6

(3) Topic Filtering Rules 3

Using rule 3 on top of rule 2, the forward themes for all themes in 2014 were calculated,

and the results are shown in Table 6. For example, calculate the similarity value between all themes in 2013 and forward theme 7, according to the highest to lowest sorting to get the highest value of 13 years of theme 1 and the forward theme 7 of the sorting position of the 6th place. The predecessor topic of the middle 5 topics is not forward topic 7, so the association of topic 7 with the middle 5 topics is removed.

Table 6. Relationship between themes and forward themes in 2014

2014 theme	prior topic (2013 theme)	similarity	Sort Maximum Topics	similarity	Is the rule valid
Topic 1	Topic 4	0.9216	11	0.9490	no
Topic 2	Topic 5	0.8287	6	0.8945	no
Topic 3	Topic 4	0.9127	11	0.9490	no
Topic 4	Topic 1	0.9261	10	0.9339	no
Topic 5	Topic 3	0.8736	1	08937	no
Topic 6	Topic 5	0.8945	6	0.8945	no
Topic 7	Topic 7	0.8407	1	0.8937	Topic 7 < x < Topic 1
Topic 8	Topic 4	0.9115	11	0.9490	no
Topic 9	Topic 6	0.8985	13	0.9186	Topic 9 < x < Topic 13
Topic 10	Topic 1	0.9339	10	0.9339	no
Topic 11	Topic 4	0.9473	11	0.9490	no
Topic 12	Topic 2	0.8956	1	0.9126	no
Topic 13	Topic 6	0.9186	13	0.9186	no
Topic 14	Topic 1	0.8571	10	0.9339	Topic 4 < x < Topic 10
Topic 15	Topic 1	08943	10	0.9339	no

To better demonstrate the flow and evolution process of topics in each period, the mulberry diagram is used to draw the topic migration degree filtered by topic filtering rules, as shown in Figure 7.

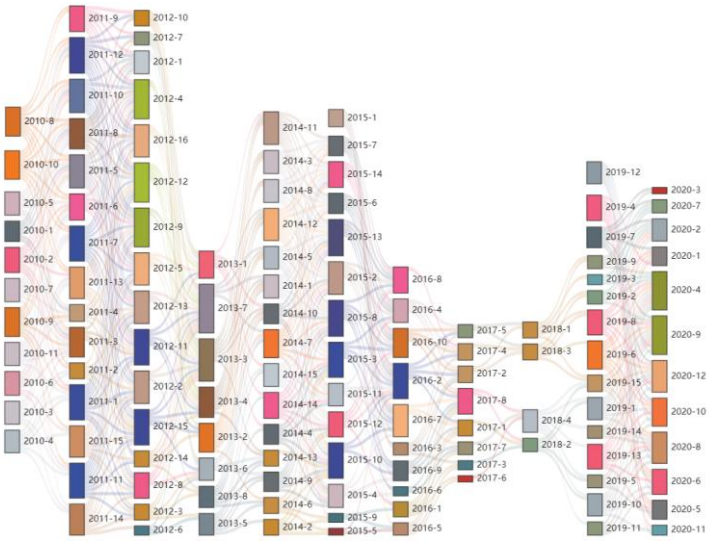

Fig. 7. Topic migration evolution

The hottest topic in the figure is selected for analysis, and the topic related to rock mass is one of the hottest topics for geologists. The two main lines of migration-evolution relationships for the rock-related topic are:

Topic2010-1→Topic2011-4→Topic2012-12→Topic2013-2→Topic2014-1→Topic2015-13→Topic2016-4→Topic2017-5→Topic2018-3→Topic2019-2→Topic2020-2

Topic2010-1→Topic2011-4→Topic2012-13→Topic2013-4→Topic2014-11→Topic2015-13→Topic2016-4→Topic2017-5→Topic2018-3→Topic2019-2→Topic2020-2

It can be seen that the theme of rock mass spread in 2012, which was expanded from one evolution relationship to two evolution paths. The first evolution path: In 2012, isotopes were added to the originally studied rock masses to study the principles of mineralization. By 2014 the element types of various ore bodies were studied. The second evolution path: various chemical elements were added to the studied rock masses, the structure of the rock masses was studied in depth in 2013, and the time of formation of the rock masses was studied in 2014. In 2015, the two evolutionary paths merged again and became one evolutionary path.

5 Conclusions

In this paper, we propose the BERT-LDA topic mining evolution model to address the problems of the traditional topic mining evolution model such as the lack of contextual semantic information, the unknown threshold of topic association screening, and edge text interference. Firstly, a sentence-transformer is used to remove the interference of edge text to topic mining, and then the BERT-LDA model is used to mine deep semantic topics. Finally, topic association filtering rules are used to remove invalid associations, and the evolution analysis is carried out from the two aspects of topic heat and topic intensity. The final experimental results show that the model effectively filters the interference of edge text on topic mining, as well as better solves the problems of contextual semantic information and topic continuity between different periods in topic mining. The effectiveness of the model is improved, and the evolution pattern of geological journals in

each period can be analyzed effectively.

Acknowledgments

This research was supported by the industry-university research innovation fund of Chinese universities (2020ITA03003) and the Informatization Project of the Fujian Education Department.

References

1. H. Qiu, B. Shao. “Research on Identification Methods of Scientific Research Hotspots under Multi-source Data,” Library and Information Service, 2020, 64(5): 78.
2. Xu Xiaoyang, Zheng Yanning, Liu Zhihui. Study on the Method of Identifying Research Fronts Based on Scientific Papers and Patents [J]. Library and Information Service, 2016, 60(24): 97.
3. Yu G, Wang M Y, Yu D R. Characterizing knowledge diffusion of nanoscience & nanotechnology by citation analysis[J]. Scientometrics, 2010, 84(1): 81-97.
4. Hou Jianhua, Wang Zhongyu. The Measurement of Knowledge Flow in Research Subject with a Empirical Analysis——Taking H-index Study as an Example [J]. Library and Information Service, 2017, 61(10): 87-93.
5. Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks [J]. arXiv preprint arXiv:1908.10084, 2019.
6. Chen Honglin, Wei Ruibin, Zhang Wei, et al. Research on Domestic Text Sentiment Analysis Based on Co-word Analysis [J]. Journal of Modern Information, 2019, 39(6): 91-101.
7. Shang Pu. On the Operation Mechanism of National Intelligence Work Statistics and Analysis of Word Frequency Based on "National Intelligence Law of the People's Republic of China" [J]. Journal of Intelligence, 2020, 39(2): 5-10.
8. Feng Guohe, Kong Yongxin. Subject Hotspot Research Based on Word Frequency Analysis of Time-Weighted Keywords [J]. Journal of the China Society for Scientific and Technical Information, 2020, 39(1): 100-110.
9. Guan Peng, Wang Yuefen. Dynamic Analysis of Authors' Research Interests in Disciplinary Field Life Cycle [J]. Library and Information Service, 2016, 60(19): 116.
10. Wang Jie, Chen Zhigang, Liu Jialing, et al. Privacy Behavior Mining Technology for Cloud Computing Based on Clustering [J]. Computer Engineering and Applications, 2020, 56(5): 80-84.
11. Li G, Zhu X, Wang J, et al. Using lda model to quantify and visualize textual financial stability report [J]. Procedia computer science, 2017, 122: 370-376.
12. Wang Yajing, Guo Qiang, Deng Chunyan, et al. Research on User Traits Predicting Based on LDA Topic Model [J]. Complex Systems and Complexity Science, 2020, 17(4): 9-15.
13. Zhang Xiuhua, Yun Hongyan, He Ying, et al. Chinese News Event Detection and Theme Extraction Based on Convolution Neural Network and K-means [J]. Science Technology and Engineering, 2020.
14. Ruan Guangce, Xia Lei. Hot Topic Detection in Journal Papers Based on Doc2Vec [J]. Information Studies: Theory & Application, 2019, 42(4): 107.
15. Hu Jiming, Qian Wei, Li Yuwei, et al. Topic Mining and Structured Parse of Policy

Text Based on LDA2Vec [J]. Information Science, 2021, 39(10): 11.

16. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv:1810.04805, 2018.
17. Wang Chunxiu, Ran Meili. Theory Foundation Discussion about Quantitative Analysis of Subjects Theme Evaluation [J]. Journal of Modern Information, 2008, 28(6): 48-50.
18. Liu Ziqiang, Wang Xiaoyue, Bai Rujiang. Research on Visualization Analysis Method of Discipline Topics Evolution from the Perspective of Multi-Dimensions: A Case Study of the Big Data in the Field of Library and Information Science in China [J]. Journal of Library Science in China, 2016, 42(6): 67-84.
19. Chen C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature [J]. Journal of the American Society for information Science and Technology, 2006, 57(3): 359-377.
20. Cobo M J, López-Herrera A G, Herrera-Viedma E, et al. SciMAT: A new science mapping analysis software tool [J]. Journal of the American Society for Information Science and Technology, 2012, 63(8): 1609-1630.
21. Wang Xiaoguang, Cheng Qikai. Analysis on Evolution of Research Topics in a Discipline Based on NEViewer [J]. Journal of the China Society for Scientific and Technical Information, 2013, 32(9): 900-911.
22. Wu Jiang, Liu Guanjun, Hu Xian. An Overview of Online Medical and Health Research: Hot Topics, Theme Evolution and Research Content [J]. Data Analysis and Knowledge Discovery, 2019 (4): 2-12.
23. Niu Li, Du Lihua. Research on the Construction and Evolution Analysis of the Subject Theme Network of Domestic Archives [J]. Archives Science Study, 2020, 34(2): 1
24. Li Yue. Research on the Evolution of Public Policy Topics in Emergent Public Health Events——Taking the Official WeChat of the National Central City as an Example [J]. Journal of Intelligence, 2020, 39(09):143-149.
25. Zhu Guang, Liu Lei, Li Fengjing. Research on Topic Relation and Prediction Based on LDA and LSTM——A Case Study of Privacy Research [J]. Modern Information, 2020.
26. Yan Duanwu, Su Qiong, Zhang Xinyu. Research on Frontier Detection in Scientific Field Based on Sequential Topic Association Evolution [J]. Information Studies: Theory & Application, 2019, 42(7): 144.
27. Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of machine Learning Research, 2003, 3(Jan): 993-1022.
28. Blei D M. Probabilistic topic models [J]. Communications of the ACM, 2012, 55(4): 77-84.