2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Retrieval and Ranking of Scientific Documents Based on LSB and TOPSIS

Peng Jin(✉)

School of Cyber Security and Computer, Hebei University, Baoding 071002, China
pengjin_hbu@126.com

Abstract. Fusing mathematical expressions, related text features, and document attributes is vital for improving the retrieval and ranking performance of scientific documents based on mathematical information. However, the specificity of mathematical expressions and their contextual relationships, as well as the diversity of document attributes, they are difficult to be fully utilized in existing models. To address these issues, a retrieval and ranking model of scientific documents based on LSB (LDA-SBERT) and TOPSIS (Technique for Order Preference by Similarity to an Ideal Solution) is proposed. Firstly, the mathematical expressions are analyzed using a symbol-level multidimensional parsing algorithm, and hesitant fuzzy sets is introduced to calculate the similarity of mathematical expressions. Then, the LSB model is used to analyze the text surrounding mathematical expressions, extract the contextual features, and calculate text similarity accordingly. By integrating the similarity of mathematical expressions and text, the preliminary retrieval results are obtained. Finally, the document attribute set is constructed, the TOPSIS is used to calculate the influence weight of documents, and weight it to the preliminary results to achieve more precise and influential ranking of scientific documents. Experimental results show that the average MAP_10 is 85.5% and the average NDCG@10 is 87.8%.

Keywords: Scientific document retrieval; document ranking; mathematical expressions; hesitant fuzzy sets; LSB; TOPSIS.

1 Introduction

In the process of scientific research and learning, it is crucial to obtain the required information quickly and accurately facing the vast and complex scientific documents. Traditional full-text document retrieval methods struggle to fully analyze and capture the deep meaning of formal content when retrieving documents containing a large number of mathematical formulas and other formalized information. There is a need to integrate mathematical expression information [8][10] to enhance the performance of scientific documents.

However, the unique two-dimensional structure of mathematical formulas, diverse symbol systems, and rich semantic information limit the effectiveness of scientific document retrieval by mechanical structural matching or semantic reasoning methods. The same mathematical expressions may have different meanings in different scenarios. and

mathematical expressions in scientific documents are closely related to their context, which often explains the application scenarios and specific meanings of the mathematical expressions. Fusing mathematical expressions and their contextual information for comprehensive retrieval [3][12] can improve the accuracy of document retrieval.

In addition, the retrieval of scientific documents should not only consider the content of the documents but also pay attention to the document's publication time, the number of citations, the situation of funding support, and other ontological attributes, which can make the retrieval and ranking results more meet the needs of users. Therefore, integrating multi-dimensional features such as mathematical expressions, text, and document attributes is an inevitable trend for achieving efficient retrieval and ranking.

Given the above problems, a scientific documents method based on LSB and TOPSIS [1] is proposed. Firstly, it parses the structure of mathematical expressions and constructs membership functions of symbol attributes to calculate the similarity of mathematical expressions. Then, the contextual features of the mathematical expression context are extracted for text matching. The preliminary retrieval results are obtained by synthesizing the expressions and text similarity. Finally, it calculates the influence weight of documents based on scientific document attributes, and weights the preliminary retrieval results to improve the accuracy of the retrieval results and the relevance of the ranking results. The contributions of this paper are as follows:

(1) Using HFS (Hesitant Fuzzy Sets) [17][19] to calculate the similarity of mathematical expressions. The method parses the mathematical multi-dimensional features of expressions at the symbol level, and use the hesitancy of fuzzy sets to calculate similarity.

(2) Constructing an LSB model to extract contextual features of mathematical expressions context. The LSB model can extract topic features and semantic features, and fuse them into contextual features to enhance feature representation, thereby calculating text similarity.

(3) Proposing a scientific document ranking method based on TOPSIS. This method introduces the ontological attributes of scientific documents to calculate the influence weight of documents, and the preliminary retrieval results are weighted to obtain more accurate and influential ranking results of scientific documents.

2 Related Work

Merely depending on text-based retrieval methods for scientific documents frequently fails to satisfy the retrieval requirements, researchers have integrated mathematical information retrieval techniques into the field of scientific document retrieval. Zhong and Zanibbi [20] used OPT node path indexing and provided K maximum common subexpressions for actual matching. They evaluated the similarity of matching OPT subtrees by calculating nodes corresponding to visible symbols and weighted operators below operands. Wang and Tian [18] proposed the CLFE model for retrieving mathematical expressions, which uses unsupervised contrastive learning techniques to embed LaTex, PMML, and CMML into the same vector space, learning the potential relationships between them to retrieve. Li et al. [6] proposed a multimodal mathematical

expression retrieval model that can utilize formulas in both image and text modes for retrieval. The retrieval list is generated through an interactive ranking fusion strategy.

Scientific document retrieval not only includes the need to retrieve mathematical formulas but also involves the retrieval of text and other content. Peng et al. [12] proposed a MathBERT pre-training model, which jointly trains mathematical expressions and their context for mathematical information retrieval. Dadure et al. [4] extracted mathematical formulas and their context in Presentation MathML format, embedded and indexed in the form of binary vectors. Pathak et al. [11] proposed a context-guided method, constructing a knowledge base containing context-formula pairs. Then, they used the Apache Lucene indexer to index the context in the knowledge base.

The multidimensional features of mathematical expressions and their contextual information should be taken into account. In this study, the retrieval of scientific documents comprehensively utilizes expressions and their contextual information.

3 Proposed Method

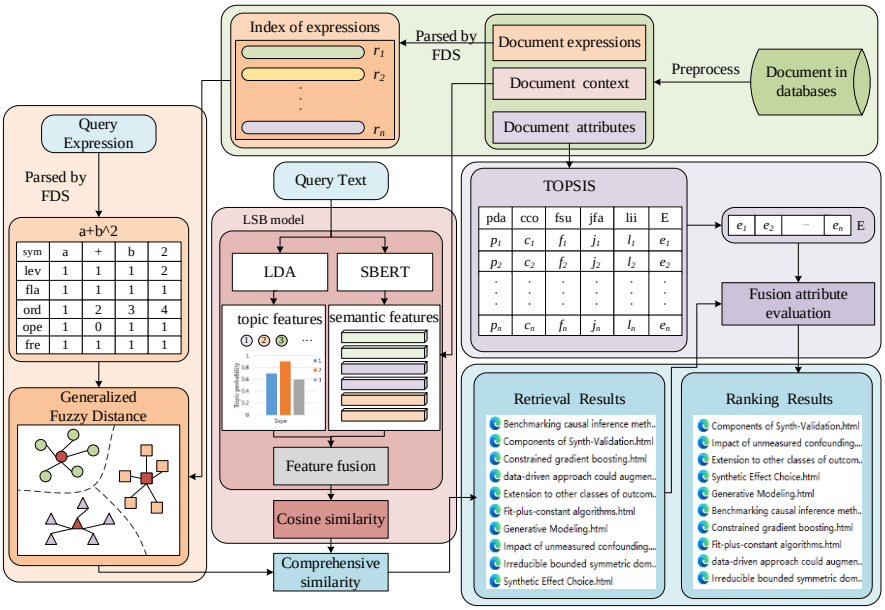

Fig. 1. Framework of scientific document retrieval and ranking model based on LSB and TOPSIS.

Scientific documents retrieval and ranking model based on LSB and TOPSIS is shown in Fig. 1. In the mathematical expression matching module, the FDS (Formula Description Structure) algorithm [16] is used to parse the structure of mathematical expressions, and then the hesitant fuzzy similarity of the mathematical expressions is calculated. In the text matching module, the LSB model is used to extract the contextual features of mathematical expressions context, and calculate the cosine similarity with

the query text features, the preliminary retrieval results are obtained by integrating the similarity of mathematical expressions and context. In the document ranking module, the TOPSIS method is used to comprehensively evaluate document attributes, obtain document influence weights, and weight the preliminary retrieval results to get the final accurate and influential ranking results.

3.1 Mathematical Expression Matching

In this section, the FDS algorithm is used to parse the expression and define the symbol membership degree function. We calculate the similarity between the query expression and the mathematical expressions in the document using the generalized hesitant fuzzy distance.

Hesitant fuzzy sets are a way to fully characterize the original information by using multiple membership degrees based on fuzzy information. The FDS algorithm can parse mathematical expressions at the symbol level to obtain five attributes for each symbol: baseline position level, spatial position flag, order of appearance, whether it is an operand operator, and frequency of appearance. Based on these attributes, membership functions are constructed, thereby forming the hesitant fuzzy features. Membership functions are as shown in Table 1.

Table 1. Definition of Membership Functions.

Membership Functions	Description
$\mathrm{h}_{\mathrm{lev}}(S_{d\_r}, S_{q\_t}) = \exp\left(-\frac{\left\lvert level_{S_{d\_r}} - level_{S_{q\_t}}\right\rvert}{level_{S_{d\_r}}}\right)$	S_{d_r}, S_{q_t} represent the r-th symbol in the storage expression and the t-th symbol in the query expression, respectively.
$\mathrm{h}_{\mathrm{fla}}(S_{d\_r}, S_{q\_t}) = \{f, \mathrm{flag}(d,q)\}$	$\mathrm{flag}(d,q)$ represent the spatial position between S_{d_r} and S_{q_t}. If they are the same, assign a value of 1; otherwise, set it to 0.。
$\mathrm{h}_{\mathrm{ord}}(S_{d\_r}, S_{q\_t}) = \exp\left(-\frac{\left\lvert ord_{S_{d\_r}} - ord_{S_{q\_t}}\right\rvert}{\sigma}\right)$	σ is a balancing factor that ensures the value of h_{ord} falls within the range of 0–1.
$\mathrm{h}_{\mathrm{ope}}(S_{d\_r}, S_{q\_t}) = \{p, p_{d\_r}\}$	When p_{d_r} represent an operator, the value of p is 1; otherwise, it is 0.
$\mathrm{h}_{\mathrm{fre}}(S_{d\_r}, S_{q\_t}) = \exp^{-\left(\frac{\left\lvert fre_{S_{d\_r}} - fre_{S_{q\_t}}\right\rvert}{\sigma}\right)^2}$	σ is a balancing factor that ensures the value of h_{fre} falls within the range of 0–1。

The generalized hesitant fuzzy distance formula for calculating the similarity of mathematical expressions is as follows:

$$\mathrm{sim}_{\mathrm{f}}(Q,D) = 1 - \left[\frac{1}{5}\sum_{i=1}^{n}\left(\frac{1}{l_{x_i}}\sum_{j=1}^{l_{x_i}}\left|h_{\mathrm{Q}}^{\sigma(j)}(x_i) - h_{\mathrm{D}}^{\sigma(j)}(x_i)\right|^{\lambda}\right)\right]^{\frac{1}{\lambda}} \tag{1}$$

Where Q represents the query mathematical expression, D represents the mathematical expression in the document, l_{x_i} is the number of evaluation values, $h_Q(x_i)$ is the HFS of query expressions, and $h_D(x_i)$ is the HFS of candidate expressions. $h_Q^{\sigma(j)}(x_i)$ is the j-th element in $h_Q(x_i)$, and $h_D^{\sigma(j)}(x_i)$ is the j-th element in $h_D(x_i)$.

3.2 Text Matching

The LDA [2] demonstrates significant advantages in topic extraction, while SBERT [13] also possesses outstanding strengths in semantic extraction. The LSB model achieves precise matching of text similarity by combining the topic modeling capabilities of LDA with the deep text embedding technology of SBERT. The construction process of the LSB model begins by inputting the preprocessed mathematical expressions contexts into the LDA model and the SBERT model separately. Next, the topic vectors output by the LDA model and the semantic vectors output by the SBERT model are connected using an autoencoder. The connected vectors are called contextual vector that enhances the expressive ability of the features The flowchart is shown in Fig. 2.

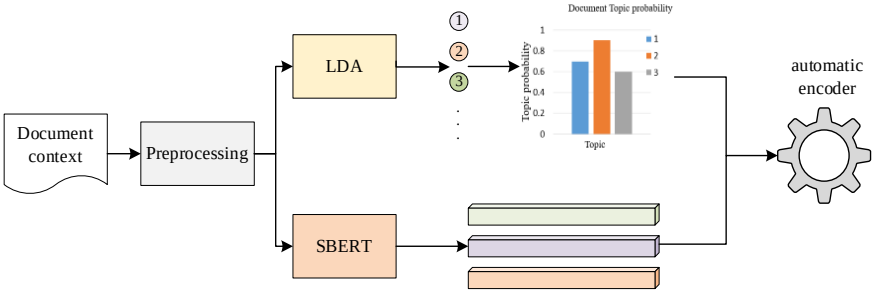

Fig. 2. LSB Framework.

Topic Feature Vectors Extracting. The topic features of the mathematical expression context are extracted by the LDA model, The LDA model is an unsupervised probabilistic topic model that reveals hidden topics in text data by capturing the distribution patterns of words in documents, and establishes probabilistic relationships between words and topics, as well as between topics and documents in a multi-layered structure. For the k-th topic, let w represent the random variable for words, then the topic probability distribution for the m-th document and the word distribution for the k-th topic satisfy the following formula:

$$p(z_m \mid \alpha) = \int p(z_m \mid \theta_m)\, p(\theta_m \mid \alpha)\, d\theta_m \tag{2}$$

$$p(w_k \mid \beta) = \int p(w_k \mid \varphi_k)\, p(\varphi_k \mid \beta)\, d\varphi_k \tag{3}$$

Where α and β are the hyperparameters of the distribution, and θ_m and φ_k are random variables. The generation process of LDA is as follows:

$$p(z_m, w_m, \theta_m, \varphi \mid \alpha, \beta) = \prod_{n=1}^{N_m} p\left(w_{m,n} \mid \varphi_{z_{m,n}}\right) p\left(z_{m,n} \mid \theta_m\right) p(\theta_m \mid \alpha)\, p(\varphi \mid \beta) \quad (4)$$

Where $z_{m,n}$ is the topic corresponding to the n-th word, $\varphi_{z_{m,n}}$ is the word distribution corresponding to $z_{m,n}$, $w_{m,n}$ is the word sampled from $\varphi_{z_{m,n}}$.

Semantic Feature Vectors Extracting. The semantic features of the mathematical expression context are extracted by the SBERT model. SBERT uses twin networks and triplet networks to capture the semantic information of sentences and update the weight parameters; it introduces pooling operations to obtain fixed-length embedded sentence vectors. As shown in Fig. 3 of the SBERT structure, the subnetwork of SBERT uses two BERT models with shared parameters.

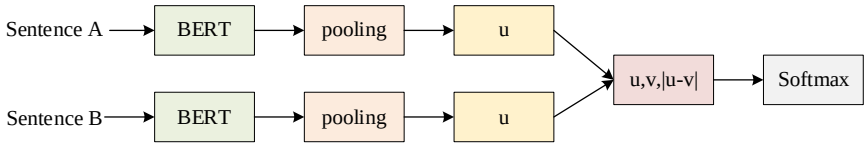

Fig. 3. Classification objective function of SBERT.

When calculating the similarity between query text and expression context, query text and expression context are encoded using BERT separately. No matter whether query text and formula context are equal in length or not, after passing through the pooling layer, they will obtain equal-length sentence vectors *q* and *f*.

Feature Fusion. After obtaining the LDA topic features vector and the SBERT semantic features vector, an autoencoder is used to concatenate them into contextual vectors. The pseudocode for the autoencoder is as follows:

Algorithm 1

Input：LDA topic features vector and SBERT semantic features vector
output：contextual vectors

```
lda_vec = get_lda_features(text)
sbert_vec = get_sbert_features(text)
combined_vec = Contact([lda_vec, sbert_vec])
input = lda_dim + sbert_dim
autoencoder = Autoencoder(input, latent_dim, activation)
    encoder_input = Input(shape=(input_dim))
    encoded = Dense(latent_dim, activation = self.activation)
    decoder_input = Input(shape=(latent_dim))
    decoded = Dense(input, activation = self.activation )
contextual_vector = autoencoder.encode(combined_vec)
```

Retrieval Results Calculating. Text similarity is calculated using cosine similarity by mathematical expression contextual features and query text features. By considering

both the similarity of the mathematical expressions and the similarity of the contextual text, the preliminary retrieval results of scientific documents are obtained.

3.3 Ranking with Influence Weights

To further enhance the accuracy and practicality of retrieval results, we comprehensively analyze the situation of Chinese and English scientific documents. We introduced multiple ontological attributes, including publication time, citation count, funding support, journal impact factor, and journal indexing indicators. Calculating the influence of the document based on the above attributes, we achieved a multi-dimensional optimization of the ranking results.

In the ranking mechanism, the document ranking is regarded as the scheme to be evaluated, and the document attributes are considered as evaluation indicators. Using the TOPSIS to evaluate the document attributes, we can obtain the document influence weights. By combining the preliminary retrieval results with the document influence weights for ranking, ranking results of more influential scientific documents can be provided.

Document Attribute Set Constructing. This paper constructs a document attributes set C = {pda, cco, fsu, jfa, lii}. The specific attributes definitions are shown in Table 2.

Table 2. Document Attribute Definitions and Assignments.

Attribute	Definition	Assignment
pda	publication date	The publication date of the document
cco	citation count	The number of citations the document
fsu	funding support	The value range is [1,5] based on the quantity and classification of funding support.
jfa	journal impact factor	The impact factor value of the journal in which the document was published
lii	journal indexing indicators	The value range is [1,5] based on the professional recognition in the relevant field of the index database, as well as the quantity and classification of the index databases.

TOPSIS. The method is a multi-attribute decision-making method that ranks alternatives by calculating relative closeness to the ideal solution and the negative ideal solution. TOPSIS can make full use of the original data information, and there are no special requirements for the evaluation data, it only requires consistency in the evaluation method for a single attribute. By standardizing the data, it can reflect the differences between alternatives objectively and truly. The steps to calculate the weights of scientific document attributes using the TOPSIS method are as follows.

(1) Construct the Evaluation Matrix

Suppose there are n documents in the preliminary search results. Extract the attribute information of each document and represent it as an evaluation matrix:

$$A = \left(x_{ij}\right)_{n*5} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \\ x_{21} & x_{22} & x_{23} & x_{24} & x_{25} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & x_{n5} \end{pmatrix} \quad (5)$$

where x_{ij} denotes the evaluation score of the j-th attribute in the i-th article. $\left(i = 1, 2, \cdots, n; j = 1, 2, 3, 4, 5\right)$

(2) Standardization Matrix

Different attributes often have different scales. To eliminate the differences among attributes, they are subjected to standardization processing. The standardization calculation method is as follows:

$$Z_{ij} = \frac{x_{ij} - \min x_j}{\max x_j - \min x_j} \quad (6)$$

(3) Solving for the Positive and the Negative Ideal Solution

Calculate the positive ideal solution and the negative ideal solution for the five indicators, respectively.

$$Z^+ = \max\left(Z_{i1}, Z_{i2}, \ldots, Z_{in}\right) \quad (7)$$

$$Z^- = \min\left(Z_{i1}, Z_{i2}, \ldots, Z_{in}\right) \quad (8)$$

(4) Calculating Euclidean Distance

Calculate the closeness of each evaluation target to the best D_i^+ and worst targets D_i^-.

$$D_i^+ = \sqrt{\sum_{j=1}^{n}\left(Z^+ - Z_{ij}\right)^2} \quad (9)$$

$$D_i^- = \sqrt{\sum_{j=1}^{n}\left(Z^- - Z_{ij}\right)^2} \quad (10)$$

(5) Calculating the comprehensive evaluation value for each document.

$$C_i = D_i^- / (D_i^+ + D_i^-) \quad (11)$$

Calculation of Weighted Similarity. The comprehensive evaluation value is used to assign weights to the influence of scientific documents. The formula for calculating the weighted similarity of scientific documents is as follows:

$$rsim_i = C_i * sim_i \quad (12)$$

Where $rsim_i$ represents the weighted similarity, sim_i represents the preliminary retrieval similarity. The final ranking results are obtained by sorting based on the weighted similarity.

4 Experiment and Result Analysis

4.1 System Experiments

The English dataset is a free and open ArXiv dataset [14] provided on Kaggle, which contains metadata for more than 1.7 million English research papers. 12,000 English papers were selected for experimentation, from which 4,964,693 mathematical expressions were extracted. A Chinese dataset [15] was introduced, which includes 10,372 Chinese scientific documents and 121,495 mathematical expressions. To verify the effectiveness of our method, we selected 10 mathematical expressions with diverse structures and related texts from different fields. The details are shown in Table 3.

Table 3. Query Input.

No.	Query Expression	Query Text
1	$E = mc^2$	mass-energy equivalence
2	$\frac{TP}{TP + FN}$	This article uses four indicators that are commonly used in classifications
3	$\sum_{i=1}^{n} a_j^2 = 1$	The summation formula also has applications in lattice theory and algebraic geometry.
4	$x + y$	consider the communication complexity of computing an approximation n
5	$\prod_{i=1}^{N} p(x_i \mid C_k)$	Suppose there are K categories of data, Naive Bayes uses the following formula to determine the category to which a data x belongs
6	$SU(2)_L \times SU(2)_R$	PNGB's arising from spontaneously broken chiral symmetry Georgi
7	$\sqrt{x^2 + y^2}$	Combined with the length of the vector, the formula for defining the fuzzy vector is as follows
8	$\sqrt{ab} \le \frac{a+b}{2}$	The arithmetic mean of two nonnegative real numbers is greater than or equal to their geometric mean
9	$\int_{t-1}^{t} \sigma_s^2 ds$	We can get the positive and negative jump components given by the following formula
10	$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}$	The Taylor function of a real or complex-valued function

To verify the performance of the model, this paper selects two commonly used indicators: MAP (Mean Average Precision) and NDCG (Normalized Discounted Cumulative Gain). Table 4 shows the average MAP_k values for the retrieval results of mathematical expressions. Fig. 4 illustrates NDCG@10 of ranking across different queries.

Table 4. MAP_k of retrieval results under different datasets.

Dataset	MAP_5	MAP_10	MAP_20
English dataset	0.922	0.891	0.768
Chinese dataset	0.897	0.864	0.798

Fig. 4. NDCG@10 of ranking across different queries.

4.2 Ablation Experiments

Different experimental models for scientific document retrieval were constructed according to different methods. Experiment 1 used the FH (FDS & HFS) model of mathematical expression retrieval to retrieve scientific documents. The LSB model was applied to Experiment 2. Experiment 3 employed the FHLS (FH & LSB) model to use mathematical expressions and contexts for co-retrieval. Experiment 4 involved the FHLST (FHLS & TOPSIS) model, which contained all the above models and introduced attributes of scientific documents. Table 5 and Fig. 5 present the average MAP_k and average NDCG@k of the ablation experiment under English and Chinese datasets separately.

Table 5. MAP_k of retrieval results under different datasets.

Method	MAP_5		MAP_10	
	English	Chinese	English	Chinese
FH	0.856	0.806	0.813	0.754
LSB	0.841	0.810	0.809	0.737
FHLS	0.896	0.828	0.833	0.761
FHLST	0.898	0.819	0.855	0.752

It can be observed that both mathematical formulas and text can retrieve scientific documents. The retrieval performance of scientific documents is improved by integrating expressions with text. The proposed model exhibits superior performance in terms of NDCG@5. Moreover, after incorporating the document influence weight, there is an increase in the NDCG@10 value, indicating that the introduction of document attributes can further meet user needs in retrieval and ranking results.

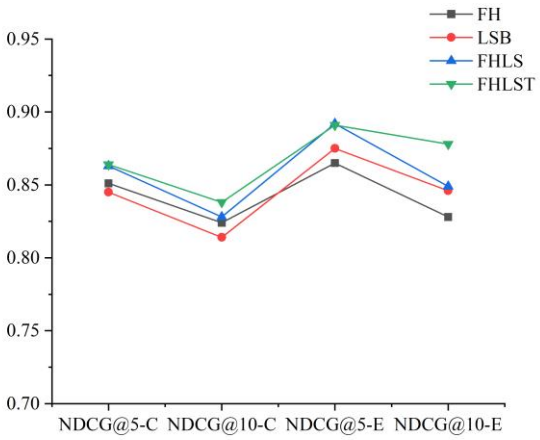

Fig. 5. NDCG@k of ranking under Chinese and English datasets.

4.3 Comparative Experiments

To verify the effectiveness of the proposed method, the following models were selected for comparative experiments: The CLFE model learns the underlying structure and content information of formulas through contrastive learning and generates formula embeddings for formula retrieval. Tangent-CFT [9] combines SLT trees with OPT trees to represent mathematical expressions, and uses the FastText model to convert formulas into vector representations for retrieval. ColBert [5] is a late interaction-based ranking model that encodes text word vectors and computes the correlation between query text and documents using a late interaction approach. Roberta [7] enhances the pre-training strategy of the BERT model by using dynamic masking strategy and a byte-level vocabulary for training, thereby extracting text features and calculating similarity for retrieval.

Table 6. Comparison of experimental results for different models.

Model	MAP_5	MAP_10	NDCG_5	NDCG_10
Clfe	0.868	0.842	0.881	0.860
Tangent-CFT	0.816	0.762	0.848	0.857
ColBert	0.918	0.850	0.872	0.863
RoBERTa	0.856	0.803	0.876	0.827
Ours	0.898	0.855	0.891	0.878

Table 6 shows the comparison results using different models. The CLFE and Tangent-CFT models rely only on mathematical expressions for retrieval and do not take into account other characteristics of scientific documents, resulting in ranking of scientific documents that is not relevant to the desired topic. Both the ColBERT and RoBERTa model focus on the encoding of query text and scientific documents text, but ColBERT

uses a post-interactive method for retrieval, which makes it better in terms of Map_5 compared to the other methods. The proposed method is overall superior to the model that only uses mathematical expressions and text. This is mainly due to deep integration of formula, text and scientific documents attributes. The comparative experiments demonstrate the applicability and potential of the method proposed for retrieving scientific documents that contain a large number of specialized formulas and terminology.

A statistical analysis using the Friedman test was conducted on the ranked values of retrieval results from various methods to determine significant differences among them. As shown in Fig. 6, the horizontal axis represents the methods, and the vertical axis represents the ranked values. It can be observed from the figure that there are significant differences in the retrieval performance of the five methods. The results indicate that the ranked values of our method are relatively higher, confirming that the method we proposed has better ranking rationality in retrieval.

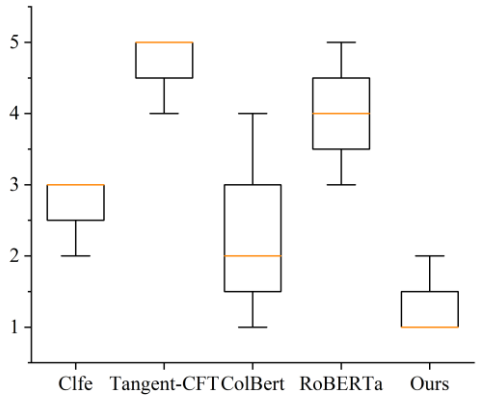

Fig. 6. The Friedman test results under different models.

5 Conclusion

This paper presents a retrieval and ranking method for scientific documents based on LSB and TOPSIS. Initially, the method parses mathematical expressions at the symbol level and introduces HFS to calculate the mathematical expressions similarity from a multi-feature perspective. Next, it constructs the LSB model to extract the contextual features of mathematical expressions context, which integrate the thematic and semantic features of the mathematical expressions context to enhance feature representativeness, thereby calculate the text similarity. Finally, the set of document attributes is constructed to calculate the influence weights of the scientific documents based on TOPSIS. The ranking results are obtained by weighting the influence weights on the retrieval results. Experimental results indicate that the method is capable of retrieving and ranking scientific documents that meet the query requirements.

In the future, we plan to focus on exploring how to effectively integrate other scientific documents elements and extracting more attributes of scientific documents to better meet user needs in the retrieval and ranking of scientific documents.

References

1. Behzadian, M., Khanmohammadi Otaghsara, S., Yazdani, M., Ignatius, J.: A state-of the-art survey of topsis applications. Expert Systems with Applications **39**(17), 13051–13069 (2012).
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Chen, C., Dai, Y., Shen, Y., Cai, J., Chen, L.: Using surrounding text of formula towards more accurate mathematical information retrieval. In: International Conference on Software Engineering and Knowledge Engineering. pp. 622–627 (2021)
4. Dadure, P., Pakray, P., Bandyopadhyay, S.: Embedding and generalization of formula with context in the retrieval of mathematical information. Journal of King Saud University-Computer and Information Sciences **34**(9), 6624–6634 (2022).
5. Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over bert. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 39–48. Association for Computing Machinery, New York, NY, USA (2020).
6. Li, R.; Wang, J.; Tian, X. A Multi-Modal Retrieval Model for Mathematical Expressions Based on ConvNeXt and Hesitant Fuzzy Set. Electronics, **12**(20): 4363 (2023)
7. Liu, Y. Roberta: A robustly optimized bert pretraining approach (2019), https://arxiv.org/abs/1907.11692.
8. Mansouri, B., Oard, D.W., Zanibbi, R.: Contextualized formula search using math abstract meaning representation. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. pp. 4329–4333. Association for Computing Machinery, New York, NY, USA (2022)
9. Mansouri, B., Rohatgi, S., Oard, D.W., Wu, J., Giles, C.L., Zanibbi, R.: Tangent-cft: An embedding model for mathematical formulas. In: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval. pp. 11–18. Association for Computing Machinery, New York, NY, USA (2019).
10. Nishizawa, G., Liu, J., Diaz, Y., Dmello, A., Zhong, W., Zanibbi, R.: Mathseer: A math-aware search interface with intuitive formula editing, reuse, and lookup. In: Advances in Information Retrieval. ECIR 2020. pp. 470–475. Springer International Publishing, Cham (2020)
11. Pathak, A., Pakray, P., Das, R.: Context guided retrieval of math formulae from scientific documents. Journal of Information and Optimization Sciences **40**(8), 1559–1574 (2019).
12. Peng, S., Yuan, K., Gao, L., Tang, Z.: Mathbert: A pre-trained model for mathematical formula understanding (2021), https://arxiv.org/abs/2105.00377
13. Reimers, N.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv: 1908.10084 (2019)
14. arXiv. org, "ArXiv dataset" Kaggle, Tech. Rep. 2023.
15. Tian, X., Wang, J.: Retrieval of scientific documents based on hfs and bert. IEEE Access **9**, 8708–8717 (2021).

16. Tian, X., Yang, S., Li, X., Yang, F.: An indexing method of mathematical expression retrieval. In: Proceedings of 2013 3rd International Conference on Computer Science and Network Technology. pp. 574–578. Dalian, China (2013).
17. Torra, V.: Hesitant fuzzy sets. International Journal of Intelligent Systems **25**(6), 529–539 (2010).
18. Wang, J., Tian, X.: Math information retrieval with contrastive learning of formula embeddings. In: Web Information Systems Engineering-WISE 2023. pp. 97–107. Springer Nature Singapore, Singapore (2023)
19. Zadeh, L.: Fuzzy sets. Information and Control **8**(3), 338–353 (1965)
20. Zhong, W., Zanibbi, R.: Structural similarity search for formulas using leaf-root paths in operator subtrees. In: Advances in Information Retrieval. ECIR 2019. pp. 116–129. Springer International Publishing, Cham (2019)