IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 4, 2020, accepted December 30, 2020, date of publication January 5, 2021, date of current version January 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3049391

Retrieval of Scientific Documents Based on HFS and BERT

XUEDONG TIAN AND JIAMENG WANG
School of Cyber Security and Computer, Hebei University, Baoding 071002, China

Corresponding author: Xuedong Tian (xuedong_tian@126.com)

This work was supported by the National Natural Science Foundation of China under Grant 61375075, the Natural Science Foundation of Hebei Province of China under Grant F2019201329, and the Key Project of the Science and Technology Research Program in University of Hebei Province of China (the Science and Technology Project of Hebei Education Department) under Grant ZD2017208.

ABSTRACT When retrieving scientific documents with mathematical expressions as the main content, both mathematical expressions and their contextual text features require consideration. However, mathematical expressions are different from texts in terms of grammar and semantics. Thus, integrating the above features and realizing scientific document retrieval is difficult. In this study, a retrieval method of scientific documents based on HFS (Hesitation Fuzzy Sets) and BERT (Bidirectional Encoder Representations from Transformer) is proposed. This method is realized through utilizing the advantages of HFS in multi-attribute decision making and BERT in context-dependent similarity calculation. By analyzing mathematical expressions and calculating the membership degree of symbolic multi-attributes, the similarity of mathematical expressions can be obtained, which can improve the accuracy of mathematical expression recall. With the extraction of the text of the expression context, BERT is used to calculate the context similarity. Then, the recalled technical documents are sorted according to the similarity of context, and the final retrieval result can be obtained. Experiments were carried out on 10,372 Chinese and 11,770 English scientific documents in the NTCIR extended data set. The average value of MAP_k($k = 10$) for the recall results of scientific documents was 74.13%. The average nDCG ($n = 10$) for the ranking of scientific documents was 86.04%.

INDEX TERMS BERT, Context, HFS, mathematical expression, scientific document retrieval.

I. INTRODUCTION

Scientific documents carry important information about scientific research and technological development. They contain a large number of mathematical expressions and texts with mathematical semantics. Retrieving scientific documents based on mathematical expressions is conducive to scientific and technological information exchange. But it is very difficult to retrieve scientific documents with mathematical expressions because mathematical expressions are characterized by a complex two-dimensional structure. To date, research on mathematical expression retrieval has achieved abundant results [1]–[7].

However, the simple matching mode of scientific retrieval, which is based merely on mathematical expressions, cannot meet the retrieval requirements of scientific documents. It is desirable to combine mathematical expressions and the meaning of their context when retrieving scientific documents. Integrating mathematical expressions and contextual text is an inevitable trend in the research of scientific document retrieval models with higher retrieval performance.

In terms of optimizing the results of sorting scientific documents based on mathematical expressions, Hussain and Khoja [8] developed a retrieval method based on the semantic information of mathematical expressions. The variables, constants and operators of expressions were replaced with unified symbols, which can help to search mathematical expressions semantically. And the weights of the semantic subtrees of mathematical expressions are assigned to make the sorting of scientific documents more reasonable. Yuan *et al.* [9] added the importance of the formula in documents and the correlation between documents while retrieving the formula, so that the result was not only the ranking of the similarity of the formula itself.

In the blend of mathematical expressions with keywords in scientific documents, Líška *et al.* [10] combined keywords with the original query of expressions and split them into subqueries. In this technique, each subquery produces an ordered list of results, each query result has its own weight. The final query result is appropriately combined with the weight to

The associate editor coordinating the review of this manuscript and approving it for publication was Gianmaria Silvello.

provide more results relevant to the original topic. Xu and Xu [11] used the formula coverage between documents to represent the formula similarity. They used the distance of feature words, which consists of the feature words in the document, to represent the text similarity. Finally, the text similarity and formula similarity are used to represent the similarity of scientific documents.

In the fusion of mathematical expressions with their context text, Pathak *et al.* [12] extracted ''context-formula'' pairs, stored this information in the knowledge base (KB), and established relevant indexes to realize the matching of mathematical text to formula. In the same year, that team developed a mathematical information retrieval system [13]. The system is divided into three modules. In the ME module, the mathematical formula is converted into a binary vector, the vector is used to calculate the similarity. The TE module evaluates the similarity of the text associated with the expression. The TME module integrates the retrieval results of the first two modules and sorts them comprehensively. Kristianto *et al.* [14] also added contextual text information while retrieving the formula of interest. They selected ten words around the formula as text information. This method improved the retrieval accuracy.

In summary, there still exist some problems in integrating mathematical expression and text information in scientific document retrieval. For example, most related achievements involve scientific documents in English. There still exist some problems to be solved in the retrieval of mathematical expressions themselves. Most of the retrieval is done for mathematical expressions in only one format (LaTeX or MathML), which cannot be unified across all formats. In terms of fusion retrieval of mathematical expressions and scientific documents, keywords extracted from scientific documents are not closely related to mathematical expressions themselves. Even if the expression and its context were combined with retrieval, the rationality and validity of the feature fusion would need to be improved.

Given the above problems, a scientific document retrieval method based on HFS (Hesitation Fuzzy Sets) [15] and BERT (Bidirectional Encoder Representations from Transformer) [16] is proposed. This system integrates mathematical expressions and their contextual text features to improve the retrieval accuracy. In this study, the format of mathematical expression retrieval is no longer restricted, both mathematical expressions in LaTeX or MathML format can be retrieved. Moreover, the retrieved data set has been expanded with the addition of Chinese scientific documents. In addition, when extracting context keywords, they are keywords that incorporate contextual information. The integration of mathematical expressions with their context makes the retrieval and ranking of scientific documents more reasonable.

II. OVERVIEW

Figure 1 shows a flow chart of the scientific documents retrieval system (dotted lines denote online query flows, and solid lines denote offline index flows). The whole process consists of three parts: Document Preprocess, Mathematical Expression Similarity and Text Similarity.

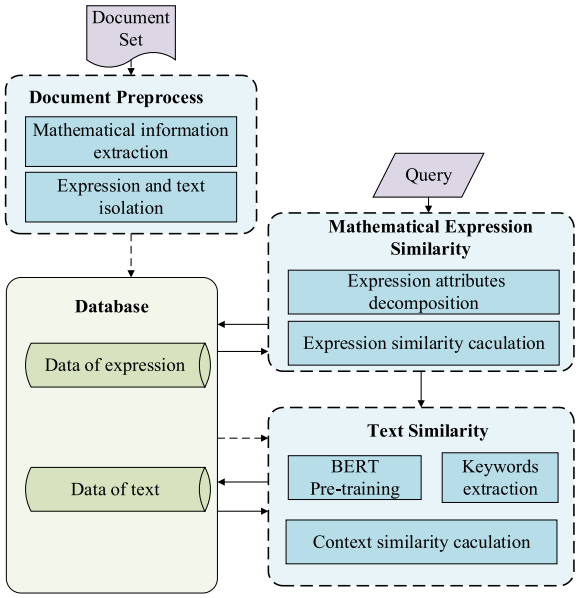

FIGURE 1. Flow chart of the scientific documents retrieval system.

The Document Preprocess module preprocesses the scientific documents in the data set. We extract the mathematical expressions in the documents and their related text parts. Then the corresponding relationship of ''mathematical expressions - text - scientific documents'' is formed.

The Mathematical Expression Similarity module is used to calculate the similarity of mathematical expressions. First, traditional FDS [17] is expanded with several attributes to each mathematical symbol. It used to decompose mathematical expressions. This approach enables the unification of LaTeX and MathML. The query expressions are disposed in the same way. Finally, the membership degree of the attribute value is calculated. The similarity of mathematical expression is obtained according to the similarity calculation formula of the HFS [18]. The ranking is returned to the Text Similarity module.

The Text Similarity module is used to calculate text similarity in the context of mathematical expressions. Relevant texts extracted from the Document Preprocess module were used to train BERT. The BERT model is used to calculate the similarity between the text corresponding to the Expression recalled in the Mathematical Expression Similarity module and the query text. Finally, the documents are sorted according to similarity.

III. SCIENTIFIC DOCUMENT PREPROCESS

The Document Preprocess module extracts mathematical expressions and their context from scientific documents and forms the list of ''mathematical expressions - context - scientific documents'' ultimately.

In scientific documents, the context of mathematical expressions contains numerous amounts of textual

TABLE 1. The context text of individual mathematical expressions in document 1 and document 2.

No.	Scientific documents (.html)	Mathematical expression	Preceding paragraphs	Following paragraphs
1	INERTIA	$p = mv$	The mass of a body determines the momentum p of the body at given velocity v; it is a proportionality factor in the formula:	The factor m is referred to as inertial mass.
2	EXPLOSIVE MATERIALS 201948-02-25-31+36	$E = 0.5CU^2$ （1）	The energy value of capacitor spark can be calculated by formula (1)://电容电火花的能量值用式 (1) 计算:	where E is the electric spark energy; C is the capacitance, U is the voltage of the charging capacitance. //式中:E 为电火花能量;C 为电容, U 为充电电容的电压。

information about them. Taking the English document "INERTIA" (as document 1) and the Chinese document "Explosive Materials 201948-02-25-31+36" (as document 2) as examples, Table 1 shows the context of the corresponding expressions. The result derived from this table shows that the context of mathematical expressions clarifies the role of the expression and explains some characters in the expression. By analyzing the content of a large number of scientific documents, the context of a mathematical expression is strongly related to the expression. Similarity is calculated using sentences and keywords extracted from the context of expressions instead of keywords extracted from scientific documents. The retrieval results obtained in this way are more in line with expectations.

First, we extract the sentences preceding and the paragraphs following the expression, and use periods and semicolons (;) to divide the sentences in the paragraph. The parameter $selists_n$ is used to represent the exp_n corresponding list of sentences. Then, dictionary A is generated, whose composition is shown as

$$\mathrm{Dic_A} = \{(exp_n, selist_n) | n \in (0, N)\} \tag{1}$$

where exp_n refers to the n-th expression and $selists_n$ refers to the list of sentences corresponding to the n-th expression.

IV. MATHEMATICAL EXPRESSION SIMILARITY

The calculation of the similarity of mathematical expressions is also the measurement of various symbols similarity and corresponding mathematical structures in the expressions. If the same mathematical symbols were to possess different attributes, their similarities would also be very different. The HFS has the advantage of processing multiple attributes. The similarity of different attribute values for a mathematical symbol is calculated, then the similarity measurement formula [18] is used to realize the retrieval of mathematical expressions.

A. HFS

HFS (Hesitating Fuzzy Sets) are mostly used in multi-attribute decision-making problems [19], [20].

Definition 1 [21]: A hesitant fuzzy set (HFS) E on $X = \{x_1, x_2, \ldots, x_n\}$ is defined in terms of a function $h_{E(x)}$ when applied to X, which returns a finite subset of [0, 1].

$$E = \left\{ < x, h_{E(x)} > | x \in X \right\} \tag{2}$$

B. EXPRESSION ATTRIBUTE DECOMPOSITION

Before the expression similarity calculation, FDS [17] is used to decompose mathematical expressions. Each mathematical symbol after decomposition has five attribute values, namely, level, flag, count, ratio and operator.

Taking the formula "$a^2 + \sqrt{b}$" as an example, the decomposition results are shown in Table 2.

TABLE 2. The result of formula "$a^2 + \sqrt{b}$" after decomposition.

Term	Level	Flag	Count	Ratio	Operator
a	0	0	1	$\frac{1}{5}$	0
2	1	2	2	$\frac{1}{5}$	0
+	0	0	3	$\frac{1}{5}$	1
$\sqrt{}$	0	0	4	$\frac{1}{5}$	1
b	1	5	5	$\frac{1}{5}$	0

C. EXPRESSION SIMILARITY CALCULATION

The expression similarity calculation algorithm is shown in Algorithm 1. Definitions 4 to 10 are the calculation methods of calling functions in the algorithm 1. Among them, definitions 4 to 8 are membership functions of the five attributes of mathematical symbols in mathematical expressions. Definition 10 is similarity calculation function.

Definition 2: E_{RE} is the query expression (LaTeX or MathML). $E_{\mathrm{SD}i}(i = 1, 2, \ldots, \mathrm{DN})$ is the mathematical

Algorithm 1 Mathematical Expression Similarity Calculation Algorithm

INPUT: $E_{\mathrm{SD}i}(i = 1, 2, \ldots, \mathrm{DN})$, E_{RE}

OUTPUT: *reslist* // A set of mathematical expressions similar to the query expression E_{RE}.

$T_{\mathrm{SD}t\_d}$ $(t\_d = 1, 2, \ldots, c\_D)$ //FDS($E_{\mathrm{SD}i}$)

$T_{\mathrm{RE}t\_r}$ $(t\_r = 1, 2, \ldots, C\_R)$ //FDS(E_{RE})

for $term_r$ in $T_{\mathrm{RE}t\_r}$:

  for $term_d$ in $T_{\mathrm{SD}t\_d}$:

    if $term_r == term_d$:

      $vec = [M_{lev}(term_r, term_d), M_{fla}(term_r, term_d), M_{cou}(term_r, term_d), M_{rad}(term_r, term_d), M_{ope}(term_r, term_d)]$

      $list_{\mathrm{med}}.add(term_d, term_d.\mathrm{id}, term_d.\mathrm{vec})$

for $term_{\mathrm{med}}$ in $list_{\mathrm{med}}$:

  for $term_{\mathrm{SD}}$ in $list_{\mathrm{SD}}$:

    if $term_{\mathrm{med}}.id == term_{\mathrm{SD}}.id$: // idin$list_{SD}$

      if $term_{\mathrm{med}}.term == term_{\mathrm{SD}}.term$: //$term$in $list_{\mathrm{SD}}$

        $list_{SD}(id, term) = \mathrm{MAX}$ (AVG($term_{\mathrm{med}}$.vec), AVG($term_{\mathrm{SD}}$.vec))

        break

    $list_{\mathrm{SD}}.add(term_{\mathrm{med}}, term_{\mathrm{med}}.id, term_{\mathrm{med}}.\mathrm{vec})$

$reslist = SIM(list_{\mathrm{SD}}, list_{\mathrm{RE}})$

RETURN *reslist*

END

expression data set, and DN is the number of mathematical expressions in the data set.

Definition 3: The term is the collective name of each mathematical symbol after the mathematical expression is decomposed. $T_{\mathrm{SD}t\_d}$ $(t\_d = 1, 2, \ldots, c\_D)$ is the t_d-th term of a mathematical expression in the data set, while c_D is the total number of terms. $T_{\mathrm{RE}t\_r}$ $(t\_r = 1, 2, \ldots, C\_R)$ is the t_r-th term of the query mathematical expression, while C_R is the total number of terms.

TABLE 3. The structure of $list_{med}$.

Term	ID of expression	Five-tuple vector
$T_{\mathrm{RE}t\_1}$	1	$(M_{lev1\_1}, M_{fla1\_1}, M_{cou1\_1}, M_{ra1\_1}, M_{ope1\_1})$
$T_{\mathrm{RE}t\_2}$	2	$(M_{lev1\_2}, M_{fla1\_2}, M_{cou1\_2}, M_{ra1\_2}, M_{ope1\_2})$
$T_{\mathrm{RE}t\_2}$	1	$(M_{lev2\_1}, M_{fla2\_1}, M_{cou2\_1}, M_{ra2\_1}, M_{ope2\_1})$
...	...	...
$T_{\mathrm{RE}t\_CR}$	n	$(M_{levCR\_n}, M_{flaCR\_n}, M_{couCR\_n}, M_{raCR\_n}, M_{opeCR\_n})$

TABLE 4. The structure of $list_{SD}$.

ID	1	...	Q
$T_{\mathrm{RE}t\_1}$	$(M_{lev1\_1}, M_{fla1\_1}, M_{cou1\_1}, M_{ra1\_1}, M_{ope1\_1})$	...	$(M_{lev1\_Q}, M_{fla1\_Q}, M_{cou1\_Q}, M_{ra1\_Q}, M_{ope1\_Q})$
$T_{\mathrm{RE}t\_2}$	$(M_{lev2\_1}, M_{fla2\_1}, M_{cou2\_1}, M_{ra2\_1}, M_{ope2\_1})$	...	(0,0,0,0,0)
...	...	...	...
$T_{\mathrm{RE}t\_CR}$	$(M_{levCR\_1}, M_{flaCR\_1}, M_{couCR\_1}, M_{raCR\_1}, M_{opeCR\_1})$	...	$(M_{levCR\_Q}, M_{flaCR\_Q}, M_{couCR\_Q}, M_{raCR\_Q}, M_{opeCR\_Q})$

TABLE 5. The structure of $list_{RE}$.

Expression	E_{RE}
$T_{\mathrm{RE}t\_1}$	(1,1,1,1,1)
$T_{\mathrm{RE}t\_2}$	(1,1,1,1,1)
...	...
$T_{\mathrm{RE}t\_CR}$	(1,1,1,1,1)

Tables 3 to 5 show the structure of the three lists in the algorithm. Table 3 shows the same terms as $T_{\mathrm{RE}t\_r}$ in the data set, as well as their expression ids and vectors of membership degrees of five groups of attributes. Table 4 is the standardization of Table 3. In Table 4, the terms that have the same expression id are unified. If an expression lacks $T_{\mathrm{RE}t\_r}$, the Five-tuple vector is set to (0, 0, 0, 0, 0). If there is the same $T_{\mathrm{RE}t\_r}$, their AVG is calculated, and the value of the Five-tuple vector is larger. Table 5 is Table 4 calculated based on E_{RE}.

Definition 4: Function $M_{lev}(T_{\mathrm{SD}t\_D}, T_{\mathrm{RE}t\_R})$ is used to calculate the membership of the level.

$$M_{lev}(T_{\mathrm{SD}t\_D}, T_{\mathrm{RE}t\_R}) = e^{-\alpha|level_{\mathrm{SD}t} - level_{\mathrm{RE}t}|} \tag{3}$$

where $level_{\mathrm{SD}t}$ and $level_{\mathrm{RE}t}$ refer to the levels of two terms and α refers to the balance factor, making $M_{lev}(T_{\mathrm{SD}t\_D}, T_{\mathrm{RE}t\_R})$ uniformly distributed in [0, 1].

Definition 5: Function $M_{fla}(T_{\mathrm{SD}t\_D}, T_{\mathrm{RE}t\_R})$ is used to calculate the membership of the flag.

$$M_{fla}(T_{\mathrm{SD}t\_D}, T_{\mathrm{RE}t\_R}) = \left\{\left(f_o, flag_{(\mathrm{SD}t, \mathrm{RE}t)}\right)\right\} \tag{4}$$

where $flag_{(\mathrm{SD}t,\mathrm{RE}t)}$ refers to the spatial position relationship of the same term in the two formulas. If $flag_{\mathrm{SD}t} = flag_{\mathrm{RE}t}$, the value of f_o is 1; otherwise, it is 0.

Definition 6: Function $M_{cou}(T_{\mathrm{SD}t\_D}, T_{\mathrm{RE}t\_R})$ is used to calculate the membership of the count.

$$M_{cou}(T_{\mathrm{SD}t\_D}, T_{\mathrm{RE}t\_R}) = e^{-(\frac{count_{\mathrm{SD}t} - count_{\mathrm{RE}t}}{\sigma})^2} \tag{5}$$

where σ is the balance factor.

Definition 7: Function $M_{rad}(T_{\mathrm{SD}t\_D}, T_{\mathrm{RE}t\_R})$ is used to calculate the membership of the ratio.

$$M_{rad}(T_{\mathrm{SD}t\_D}, T_{\mathrm{RE}t\_R}) = \frac{1}{1 + \mu \left| ra_{\mathrm{SD}t} - ra_{\mathrm{RE}t} \right|^{\upsilon}} \tag{6}$$

where μ and υ are balance factors.

Definition 8: Function $M_{ope}(T_{\mathrm{SD}t\_D}, T_{\mathrm{RE}t\_R})$ is used to calculate the membership of the operator.

$$M_{ope}(T_{\mathrm{SD}t\_D}, T_{\mathrm{RE}t\_R}) = \{(s_o, operator_{\mathrm{SD}t})\} \tag{7}$$

where $operator_{\mathrm{SD}t}$ represents whether the current term is an operator. If so, the value of s_o is 1; otherwise, it is 0.5.

Definition 9: Function AVG(*term*) is used to calculate the situation where the same term appears twice in an expression.

$$\mathrm{AVG}(term) = \frac{1}{5}\left(M_{levk\_m} + M_{flak\_m} + M_{couk\_m} M_{rak\_m} + M_{opek\_m}\right) \tag{8}$$

where k is the count of the current T_{REt_R}.

Definition 10: Function $SIM(E_{\mathrm{RE}}, E_{\mathrm{SD}i})$[21] is used to calculate the similarity of two mathematical expressions.

$$\mathrm{SIM}(E_{\mathrm{RE}}, E_{\mathrm{SD}i}) = 1 - \left\{ \frac{1}{5} \sum \left[\frac{1}{c\_R} \sum_{j=1}^{c\_R} \left| M_{j\_\mathrm{RE}} - M_{j\_\mathrm{SD}i} \right|^{\lambda} \right] \right\}^{1/\lambda} \tag{9}$$

where M_{j_RE} is the five-tuple vector value corresponding to $T_{\mathrm{RE}t\_n}$ in $list_{\mathrm{RE}}$, when $j = 1$, CR=1 in $list_{\mathrm{RE}}$. $M_{j\_\mathrm{SD}i}$ is the five-tuple vector value in $list_{\mathrm{SD}}$, and j and i correspond to CR and Q in $list_{\mathrm{SD}}$ respectively; when $j = 1$, $i = 2$, the five-tuple vector is $(M_{lev1\_2}, M_{fla1\_2}, M_{cou1\_2}, M_{ra1\_2}, M_{ope1\_2})$.$\lambda$ is the control parameter.

V. TEXT SIMILARITY

A. BERT

BERT (Bidirectional Encoder Representations from Transformer) [16] is based on a large number of unlabeled corpora for training to obtain the semantic representation of the text. The schematic diagram of the BERT model is shown in Figure 2.

The internal structure of BERT is mainly a stack of multiple layers of the Transformer Encoder [22]. BERT can solve the polysemy problem, which cannot be solved with word2vec. The same keywords may have different meanings in different contexts. When calculating text similarity, BERT fully considers the context information of the word. The vector constructed by the same word in different contexts is also different, thereby improving the accuracy of calculating text similarity.

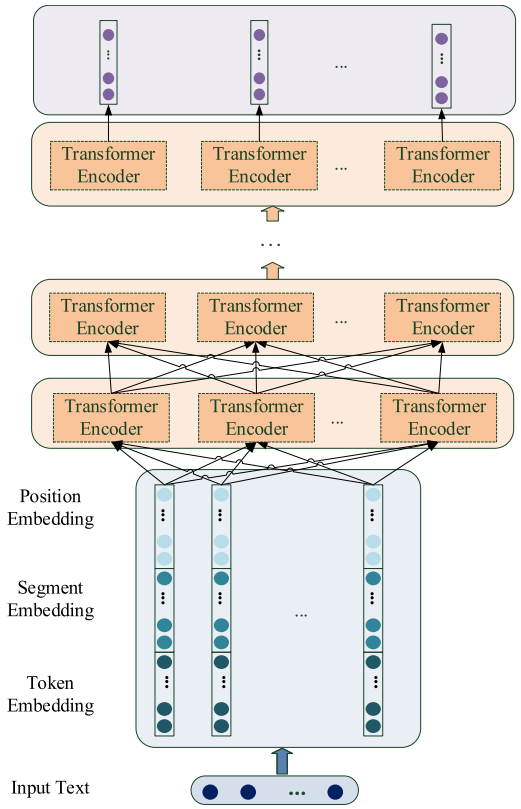

FIGURE 2. Schematic diagram of the BERT model [16].

In this study, one-third of the text in Document Text extracted by the Document preprocess module is annotated and sent to BERT for pretraining.

B. KEYWORDS EXTRACTION IN THE CONTEXT OF MATHEMATICAL EXPRESSIONS

The context of mathematical expression is the text information, which is generally closely related to the formula.

Therefore, similarity is calculated by keywords in the expression context and keywords in the query statement. Then scientific documents are finally sorted using the text similarity pairs that have been obtained.

$selists_n$ in the dictionary Dic_A are segmented, where $worists_n$ is the keyword list corresponding to exp_n. The weight of keywords is calculated as

$$W_{c,cue} = \frac{p_{c,cue}}{\sum_{j=0}^{N} \sum_{k=0}^{K} p_{k,j}} \cdot \lg \frac{N}{1 + (m_c \in N)} \tag{10}$$

where $w_{c,cue}$ refers to the weight of keyword c to the expression cue, and $p_{c,cue}$ refers to the number of times keyword c appears in the vocabulary of the expression cue. N refers to the total number of expressions, K refers to the total number of keywords in a certain expression vocabulary, and m_c refers to the number of expressions containing keyword c.

Words with a weight higher than the threshold are selected as the keywords of the current expression by setting the

Algorithm 2 Text Similarity Calculation Algorithm

INPUT: S_{RE}, $W_{SDi}(i = 1, 2, \ldots, sum)$
OUTPUT: $Simil$
$W_{RE} = jieba(S_{RE})$ //Word segmentation
$encode(S_{RE})$ // Sentence vector
for $text_r$ in W_{RE}:
  $num = \text{location}(text_r)$ // Target keywords
  summed_layers[num] //Word vector
  $simil = simil + \text{MAX}($
    $\cos ine\_\text{similarity}(\text{summed\_layers}[num], W_{SDi}))$
$simil = simil/\text{len}(W_{RE})$
RETURN $simil$
END

weight threshold. Then a dictionary B is formed. The composition of dictionary B is shown as

$$\text{Dic}_B = \{(exp_n, word_{nw}, sentence_{nw}) \mid n \in (0, \text{N}), w \in (0, \text{SUM})\} \quad (11)$$

where exp_n refers to the n-th expression. Here, $word_{nw}$ refers to the w-th keyword corresponding to the n-th expression, SUM refers to the total number of keywords in the current sentence, and $sentence_{nw}$ refers to the key sentences corresponding to $word_{nw}$.

C. SIMILARITY CALCULATION OF MATHEMATICAL EXPRESSION CONTEXT

There are many methods for calculating text similarity [23]–[25]. The current mainstream method is to use the cosine distance to calculate similarity. The closer the vector cosine distance generated by BERT, the more similar the two texts. The algorithm for calculating text similarity is shown in Algorithm 2.

Definition 11: S_{RE} is the query text. $W_{SDi}(i = 1, 2, \ldots, sum)$ is the word vector of the keywords in the trained sentence in the database, and sum is the number of keywords in the sentence.

VI. EXPERIMENTAL PROCESS AND RESULT ANALYSIS

A. EXPERIMENTAL DATA

In this study, the Chinese data set has been added to expand the data set because the public data set NTCIR contains only English documents.

There are 11770 English scientific documents containing 25,045 mathematical expressions and 10372 scientific Chinese documents containing 121495 mathematical expressions in the data set.

The details of the data set are in the appendix.

B. SYSTEM EXPERIMENT

1) RETRIEVAL RESULTS OF MATHEMATICAL EXPRESSION BASED ON HFS

A large number of experiments are conducted on different types of mathematical expressions. Then, ten expressions were chosen for statistical experiments. The expressions used in the experiment are shown in Table 6. The ten expressions contain different mathematical symbols, and the corresponding keywords involve several different professional vocabularies. The statistical experiment for these ten expressions is representative.

TABLE 6. Ten mathematical expressions in the experiment and their query text.

No.	Mathematical expressions	Query text
1	$x + y$	fundamental equation
2	$\sqrt{ab} \le \frac{a+b}{2}$	fundamental inequality
3	$\sin\alpha^2 + \cos\alpha^2 = 1$	trigonometric function
4	$y = nx^{n-1}$	differential equation
5	$f(x)$	function
6	$E = \frac{1}{2}mv^2$	theorem of kinetic energy
7	$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$	Taylor function
8	$F_g = G\frac{Mm}{r^2}$	gravity
9	$c = \frac{Q}{ct}$	specific heat capacity
10	$P = \frac{U^2}{R}$	circuit power

TABLE 7. AP_k values of mathematical expression search results.

Average Precision_k	AP_5	AP_10	AP_15	AP_20
Value	0.902	0.824	0.717	0.644

Average Precision_k refers to the average correct rate from the first correct recall to the k-th correct recall. Table 7 shows the AP_k values of different expressions under different k values. It can be seen from the table that as the value of k increases, AP_k gradually decreases, which is also related to the small number of similar expressions in the data set of some complex mathematical expressions. Because some more complex expressions have less than 20 expressions with high similarity in the data set, the value of AP_20 is slightly low.

MAP_k is used to calculate the average value of AP_k. In this study, MAP_10 = 0.7413 for the ten expressions in Table 6.

2) MATHEMATICAL EXPRESSION CONTEXT TEXT

The document "Chinese Journal of Solid Mechanics 201940-05-458-466.html" (as document 3) is used as an example. The scientific document contains multiple mathematical expressions. The context and keywords corresponding to three mathematical expressions are selected as shown in Table 8. The top three keywords are "annulus (环形)", "electromagnetic (电磁式)" and "magnet (磁铁)" in document3. Through analysis, it can be found that although the keywords of the three expressions are all related to

TABLE 8. Mathematical expressions in document 3 and their corresponding keywords.

No.	Expression in document	Keywords
1	$H=\frac{J}{4\pi\mu_0}\sum_{k=0}^{1}\sum_{i=0}^{1}\sum_{j=0}^{1}(-1)^{i+j+k}\varepsilon(U,V,W)$	magnet (磁铁) magnetic field intensity (磁场强度) magnetic flux density (磁通密度)
2	$\theta_n=\frac{n\pi}{8}(n=1,2,...,N)$	magnet (磁铁) location (所在位置) relative angles (相对角)
3	$U=n\frac{d\phi}{dt}$	capture energy (俘能) electromagnetic induction (电磁感应) Faraday (法拉第)

"magnetic", context keywords are more obviously relevant to the expressions. For instance, the No.1 expression is used to show the magnetic field intensity and magnetic flux density, the keywords in the full document do not represent these.

Therefore, the expression context keywords are more closely related to the mathematical expressions compared with the keywords in the entire scientific document.

3) RETRIEVAL RESULTS OF SCIENTIFIC DOCUMENTS BASED ON HFS AND BERT

In this study, the DCG (Discounted Cumulative Gain) is used to evaluate the sorting result, and the calculation method is shown in (11). *n*DCG is used to normalize all search results, and the calculation method is shown in (13).

$$\mathrm{DCG}=rel_1+\sum_{i=2}^{P}\frac{rel_i}{\log_2 i}(P\geq 2) \tag{12}$$

where i refers to the ranking number of the search result, rel_i refers to the classification of the i-th search result: Good, Fair, Bad, which assigned scores of 3, 2, and 1, respectively. P refers to the total number of search results.

$$n\mathrm{DCG}=\frac{DCG}{IDCG} \tag{13}$$

where IDCG refers to the ideal DCG. In other words, IDCG is the DCG calculated when the search results are sorted to the best state.

In this study, the mathematical expressions and query text in Table 6 are used for statistical experiments. The DCG obtained for different mathematical expressions is shown in Figure 3.

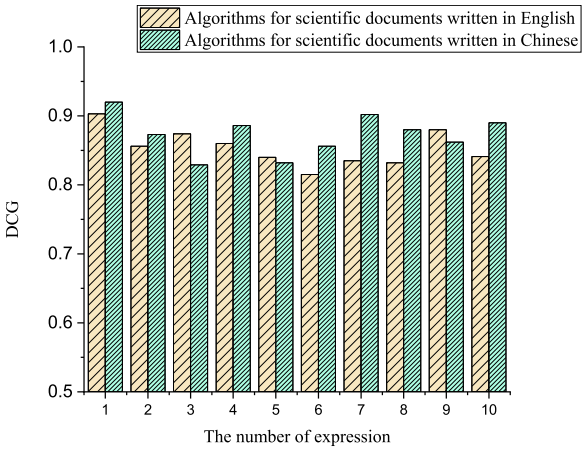

FIGURE 3. DGG value of search results under different queries.

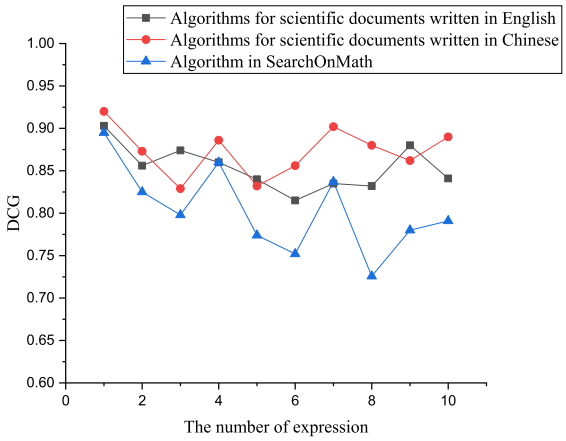

FIGURE 4. Comparison between the method in this study and the SearchOnMath method.

C. COMPARATIVE EXPERIMENT

SearchOnMath [23] is a mathematical formula retrieval tool that aims to accurately match mathematical expressions, locate scientific documents according to mathematical expressions, and then realize the retrieval of "mathematical expressions - scientific documents". However, SearchOnMath implements pure mathematical expression retrieval and does not consider the important information of the scientific document itself.

For example, expression A is used only for certain intermediate value calculations in scientific document D, and the subject of document D does not match the query subject. However, the similarity of A is high relative to query expression Q, the ranking of document D will be at the top. Obviously, there are unreasonable factors in this sort of order.

The expression retrieval results in Table 6 are used to conduct comparative experiments, and the results are shown in Figure 4.

In addition, $n\mathrm{DCG}(\mathrm{n}=10)=0.8673$ for the search results of the Chinese scientific documents. $n\mathrm{DCG}\ (n=10)=0.8535$ for the English scientific document retrieval results. The average value of $n\mathrm{DCG}$ (n = 10) is 0.8604.

TABLE 9. The data set used in the experiment of this study.

Scientific documents written in English					
Name of the dataset		Number of documents	Number of documents		
NTCIR-12_MathIR_Wikipedia_Corpus		11770	250045		
Scientific documents written in Chinese					
No.	Name of journal	No. of documents	No.	Name of journal	No. of documents
1	Explosive Materials	36	24	Computer Engineering and Science	125
2	Journal of Beijing Jianzhu University	28	25	Computer Research and Development	189
3	Chinese Journal of Chemical Engineering	69	26	Computer Applications	403
4	Journal of Chengdu Institute of Technology	27	27	Computer Applications and Software	499
5	University Maths	115	28	Computational Physics	90
6	University Physics Experiment	147	29	Journal of Building Materials	62
7	Journal of Cryophysics	21	30	Journal of Building Steel Structure Progress	13
8	Earthquake Engineering and Engineering Vibration	74	31	Journal of Building Science and Engineering	28
9	Automation of Electric Power Systems	418	32	Ship Electronic Countermeasures	24
10	Electric Automation	170	33	Journal of Jiangsu Institute of Architecture and Technology	16
11	Electronic Measurement Technology	310	34	Fuzzy Systems and Mathematics	90
12	Electronic Mechanical Engineering	23	35	Pattern Recognition and Artificial Intelligence	86
13	Electronic Technology	172	36	Journal of Heat Treatment of Thermal Materials	109
14	Chinese Journal of Luminescence	172	37	Journal of Shandong Jianzhu University	21
15	Advanced Mathematics Research	119	38	Journal of Shenyang Jianzhu University	39
16	Polymer Material Science and Technology	20	39	Journal of Mathematics Bulletin	119
17	Polymeric Materials Science and Engineering	72	40	Mathematics Learning and Research	793
18	High School Mathematics Teaching and Learning	306	41	Water Conservancy and Hydropower Technology	193
19	Public Health and Preventive Medicine	57	42	Journal of Water Resources and Civil Engineering	69
20	Chinese Journal of Solid Mechanics	15	43	Sichuan Building Research	18
21	Spectroscopy and Spectral Analysis	447	44	Statistical Research	101
22	Journal of the optical	497	45	Statistics and Information Forum	148
23	Acta Photonica Sinica	46	46	Microelectronics and Computer	203

VII. CONCLUSION

In this study, given the problems of mathematical information retrieval, a retrieval method combining mathematical expressions and contextual text is proposed.

First, the membership degree is calculated for the attribute of each mathematical symbol in the expression, and the similarity of the entire expression is obtained by using HFS. Next, BERT is used to calculate the expression context sentence vector, then the context keywords are extracted. The word vectors of the keywords are extracted from the sentence vector for similarity calculation. Finally, the scientific documents are sorted according to the text similarity.

TABLE 9. *(Continued)* **The data set used in the experiment of this study.**

47	Aero Weaponry	19	56	Journal of Xi'an University of Architecture & Technology	56
48	Journal of Engineering of Heilongjiang University	22	57	Journal of Cardiovascular Rehabilitation Medicine	122
49	Mechatronics Engineering	141	58	Journal of Vibration and Shock	713
50	Mechanical Science and Technology	177	59	Chinese Journal of Mechanical Engineering	33
51	Mechanical Design and Research	144	60	Chinese Journal of Health Statistics	210
52	Machinery & Electronics	89	61	Middle school Mathematics Teaching	134
53	Journal of Jilin Jianzhu University	31	62	Middle School Mathematics Monthly	194
54	Computer Simulation	848	63	Middle School Mathematics Magazine	180
55	Computer Engineering	460			
Total number of documents			10372		
Total number of expressions			121495		

In the future, the following ongoing studies will be carried out on scientific document retrieval based on mathematical expressions and their contextual text integration. The following tasks are considered.

(1) Improve the method of extracting contextual keywords in scientific documents, and increase the weight of some texts with mathematical semantics.

(2) Address special cases in which the text content corresponding to expressions is not in its context, or cases for which mathematical expressions corresponding to the text do not exist.

(3) Evaluate combining the algorithms of Chinese scientific documents and English scientific documents, so that regardless of the input language, the corresponding scientific documents in two languages can be retrieved.

APPENDIX

See Table 9.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 61375075), the Natural Science Foundation of Hebei Province of China (Grant No. F2019201329) and the Key Project of the Science and Technology Research Program in University of Hebei Province of China (the Science and Technology Project of Hebei Education Department) (Grant No. ZD2017208).

REFERENCES

[1] R. Deveaud, J. Mothe, M. Z. Ullah, and J.-Y. Nie, "Learning to adaptively rank document retrieval system configurations," *ACM Trans. Inf. Syst.*, vol. 37, no. 1, pp. 1–41, Jan. 2019, doi: 10.1145/3231937.

[2] K. Yamada and H. Murakami, "Mathematical expression retrieval in PDFs from the Web using mathematical term queries," in *Proc. 33rd Inter Conf. Ind., Eng. Appl. Intel Syst.*, Kitakyushu, Japan, Sep. 2020, pp. 155–161, doi: 10.1007/978-3-030-55789-8_14.

[3] P. Sojka, M. Růžika, and V. Novotný, "MIaS: Math-aware retrieval in digital mathematical libraries," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, Turin, Italy, Oct. 2018, pp. 1923–1926, doi: 10.1145/3269206.3269233.

[4] D. Stalnaker and R. Zamibbi, "Math expression retrieval using an inverted index over symbol pairs," *Doc Rec. Ret*, vol. 9402, no. 7, pp. 1–12, Feb. 2015, doi: 10.1117/12.2074084.

[5] W. Zhong, S. Rohatgi, J. Wu, C. Lee, and R. Zanibbi, "Accelerating substructure similarity search for formula retrieval," in *Proc. ECIR*, Lisbon, Portugal, Apr.2020, pp. 714–727, doi: 10.1007/978-3-030-45439-5_47.

[6] M. Schubotz, N. Meuschke, T. Hepp, H. S. Cohl, and B. Gipp, "VMEXT: A visualization tool for mathematical expression trees," in *Proc. CICM*, Edinburgh, U.K., Jul. 2017, pp. 340–355, doi: 10.1007/978-3-319-62075-6_24.

[7] R. M. Oliveira and F. B. Gonzaga, "A distributed system for searchOnMath based on the microsoft BizSpark program," in *Proc 3rd Bra Syst. Data*, Brazilian, South America, 2018, pp. 289–294.

[8] S. Hussain and S. Khoja, "Retrieval of mathematical information with syntactic and semantic structure over Web," *Jour Infor Sci Engin*, vol. 36, no. 1, pp. 75–89, 2020, doi: 10.6688/JISE.202001_36(1).0005.

[9] K. Yuan, L. Gao, Y. Wang, X. Yi, and Z. Tang, "A mathematical information retrieval system based on RankBoost," in *Proc. 16th ACM/IEEE-CS Joint Conf. Digit. Libraries*, Beijing, China, Jun. 2016, pp. 259–260.

[10] P. Sojka, and M. Růžika, "Combining text and formula queries in math information retrieval: Evaluation of Query results merging strategies," in *Proc. 1st Int. Workshop Novel Web Search Interface Syst.*, Melbourne, VIC, Australia, Oct. 2015, pp. 7–9, doi: 10.1145/2810355.2810359.

[11] J. M. Xu and C. Y. Xu, "Computing similarity of scientific socuments based on texts and formulas," *Data Know. Discovery*, vol. 22, no. 10, pp. 103–109, 2018, doi: 10.11925/infotech.2096-3467.2018.0211.

[12] A. Pathak, P. Pakray, and R. Das, "Context guided retrieval of math formulae from scientific documents," *J. Inf. Optim. Sci.*, vol. 40, no. 8, pp. 1559–1574, Nov. 2019, doi: 10.1080/02522667.2019.1703255.

[13] A. Pathak, P. Pakray, and A. Gelbukh, "Binary vector transformation of math formula for mathematical information retrieval," *J. Intell. Fuzzy Syst.*, vol. 36, no. 5, pp. 4685–4695, May 2019, doi: 10.3233/JIFS-179018.

[14] G. Y. Kristianto, G. Topia, and A. Aizawa, "Utilizing dependency relationships between math expressions in math IR," *Inf. Retr. J.*, vol. 20, no. 2, pp. 132–167, Mar. 2017.

[15] V. Torra, "Hesitant fuzzy sets," *Interface J. Int. Syst.*, vol. 25, no. 6, pp. 529–539, Mar. 2010, doi: 10.1002/int.20418.

[16] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Amsterdam, The Netherlands, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[17] X. D. Tian, "A mathematical indexing method based on the hierarchical features of operators in formulae," *Adv. Eng. Res.*, vol. 119, pp. 49–52, Aug. 2017, doi: 10.2991/icacie-17.2017.11.

[18] L. N. Cai, "Interval-valued hesitant fuzzy sets and its application to decision making," M.S. thesis, Dept. Electron. Eng., Zhengzhou Univ., Zhengzhou, China, 2013.

[19] L. Zhang, Q. Li, and J. Yu, "Patent quality evaluation model based on hesitant fuzzy soft set," in *Proc. 24th Int. Conf. Ind. Eng. Man*, Changsha, China, 2018, pp. 621–628, doi: 10.1007/978-981-13-3402-3_65.
[20] C. Zhang, D. Li, Y. Zhai, and Y. Yang, "Multigranulation rough set model in hesitant fuzzy information systems and its application in person-job fit," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 4, pp. 717–729, Apr. 2019, doi: 10.1007/S13042-017-0753-X.
[21] B. Zhu and Z. Xu, "Probability-hesitant fuzzy sets and the representation of preference relations," *Technol. Econ. Develop. Econ.*, vol. 24, no. 3, pp. 1029–1040, May 2018, doi: 10.3846/20294913.2016.1266529.
[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkorcit, L. Jones, A. N. Gomeza, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, Long Beach, CA, USA, 2017, pp. 6000–6010.
[23] C. L. Wang, Y. H. Yang, F. Deng, and H. Y. Lai, "A review of text similarity approaches," *Inf. Sci*, vol. 37, no. 3, pp. 1007–7634, Mar. 2019, doi: 10.13833/j.issn.1007-7634.2019.03.026.
[24] H. T. Nguyen, P. H. Duong, and E. Cambria, "Learning short-text semantic similarity with word embeddings and external knowledge sources," *Knowl.-Based Sys*, vol. 182, no. 15, pp. 1–9, Oct. 2019, doi: 10.1016/j.knosys.2019.07.013.
[25] M. I., W. H., and H. Abdalhakim, "A hybrid model for paraphrase detection combines pros of text similarity with deep learning," *Int. J. Comput. Appl.*, vol. 178, no. 20, pp. 18–23, Jun. 2019, doi: 10.5120/IJCA2019919011.

XUEDONG TIAN received the B.S. degree from the Department of Automation Engineering, Hebei University of Technology, China, in 1984, the M.S. degree from the Department of Electronics and Information Engineering, Hebei University, Baoding, China, in 1998, and the Ph.D. degree from the College of Physics Science and Technology, Hebei University, in 2007. He is currently a Professor with the School of Cyber Security and Computer, Hebei University. His research interests include information retrieval and pattern recognition.

JIAMENG WANG was born in Shijiazhuang, Hebei, China, in 1995. She received the B.S. degree in network engineering from Hebei University, Baoding, China, in 2018, where she is currently pursuing the master's degree under the supervision of Prof. Tian. Her main research interests include intelligent image and text information retrieval.