ATLANTIS PRESS

Advances in Engineering Research, volume 119

2nd International Conference on Automatic Control and Information Engineering (ICACIE 2017)

A Mathematical Indexing Method Based on the Hierarchical Features of Operators in Formulae

Xuedong Tian
School of Computer Science and Technology
Hebei University
Baoding, China
xuedong_tian@126.com

***Abstract*—Full text search engines widely used today still have no math searching function, which brings inconvenience for people finding their scientific documents with mathematical query words. It is necessary to research and develop the theory and technology of mathematical expression retrieval. This paper proposed an index model of mathematical expressions for realizing math retrieval through analyzing the characteristics of formulae. Firstly, the FDS data was obtained from the formulae expressed in LaTeX description with recursive analysis. Then, the index features including the level and location features of operators were extracted from the FDS data of formulae. Finally, the extracted features were used to construct a feature vector for dividing formulae into several classes and the math index was constructed for the classes respectively. The experiment was carried out on 134199 formulae and the result shows its effectiveness for improving the efficiency of mathematical expression retrieval.**

Keywords—mathematical expression retrieval; index; hierarchical features; operators

I. INTRODUCTION

As an important component of information retrieval, mathematical expression indexing and matching has been researched for many years. In scientific documents, formulae play a key role for expressing mathematical meanings in formalized manner. Therefore, the formula retrieval is an essential function of a searching engine. Unfortunately, the current widely used search engines still have no math searching function, which brings inconvenience for people who study or work in scientific fields. So it is necessary for us to research and develop the theory and technology of mathematical expression retrieval for realizing obtains the required scientific documents through math keywords.

Mathematical expression retrieval has been researched for many years. Several prototype systems have been proposed with mathematical expression searching function. DLMF Search [1] is a math retrieval system designed for searching math contents in DLMF (Digital Library of Mathematical Functions). It belongs to the type of the retrieval system which employs the traditional full text searching engine to realize math retrieval through converting formulae into text strings. MathDex [2] is a math retrieval method based on Lucene. Considering the attributes of formulae, a weight value related to the complexity, level and length of formula was defined. Libbrecht and Melis[3] described the mathematical retrieval method in ActiveMath. As a web learning benchmark, ActiveMath used OMDoc language to store its contents. Its math search function focuses on the math content items. Math searching function was realized through converting math content into token and indexing them with Lucene tool. EgoMath [4] belongs to a full text searching engine which could provide math searching function in Wikipedia. MathWebSearch [5] indexed the formulae in Content MathML description and realized its math search engine based on Apache Solr – ElasticSearch. WikiMirs [6-7] is a math search engine designed for WikiPedia. It realized math retrieval through extracting terms from formulae with a series of tree transformation. LaTeXSearch [8] is provided by Springer which could realize formula retrieval in its digital library. Moreover, many math retrieval methods were proposed which carried out benefit work for realizing formula retrieval [9-11].

Because of the complexity of formulae, there are many problems to be solved in math retrieval. In this paper, focusing on the math index construction, a hierarchical feature based indexing model is designed for improving the efficiency of math indexing and matching.

II. HIERARCHICAL FEATURE EXTRACTION OF THE SYMBOLS IN MATHEMATICAL EXPRESSIONS

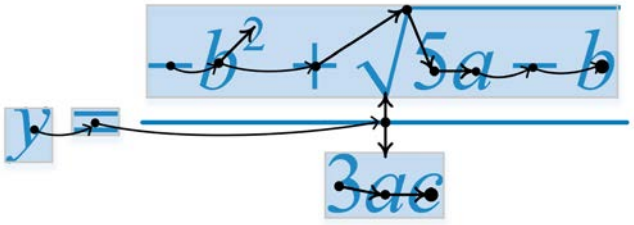
In normal text, characters are simply arranged in linear mode. The relationship between the two nearest characters is the precursor and successor for each other. And these characters are generally belonging to the same character set. However, in formulae, the symbols not only often comes from several symbol sets such as digits, English letters, Greek letters and math operators, but also are arranged in two dimensional modes such as fractions, superscripts, subscripts, radicals and so on. Figure 1 shows the difference between text and formula.

In literature [12], we defined the retrieval feature description structure of formulae called FDS (Formula Description Structure). In FDS, a symbol S_i of a formula *ME* is described as a 5-tuple array:

$$A(S_i) = (C_i, P_i, L_i, O_i, F_i) \quad (1)$$

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

(a) Linear mode of normal text



(b) The two dimensional mode of formulae

Fig. 1. Difference of the structure between text and formula

Where

- C_i is the code of S_i.
- P_i is the ordinal number of S_i in *ME*.
- L_i is the level of S_i in *ME*, which is identified by the baseline number we will illustrate later.
- O_i is the functional code of S_i. $O_i = 0$ means S_i is a operand. $O_i = 1$ means S_i is an operator.
- F_i is the relationship of S_i to its nearest prior in the higher level. Its value is from 1 to 8 respectively represents the up, superscript, right, subscript, down, inclusion, left superscript and left subscript. All the symbols in main baseline have the F_i value of 0.

The value of L_i is measured according to the baseline [10] the symbol S_i lies. In mathematical expressions, operators could be divided into two classes called one dimensional operator and two dimensional operators in the respect of the influence of the geometrical distribution of formula symbols. The former includes "+", "-", "/" and so on. The examples of the later are the symbols which could lead to the vertical distributions of formula symbols such as "-(fraction)", "superscript", "subscript", "radical". Fig. 2. shows the situation of the baselines in a formula.

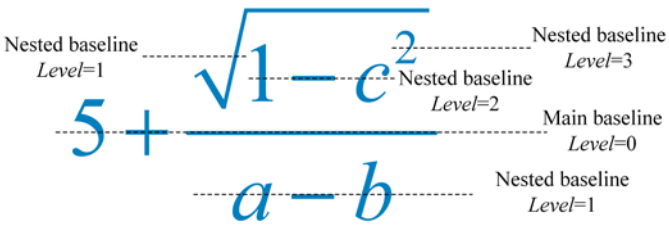


Fig. 2. Baselines in a formula

The LaTeX description of this formula is "\[5 + \frac{{\sqrt {1 - {c^2}} }}{{a - b}}\]". The one dimensional operators are "+" and "-", while the two dimensional operators are "$\frac{\square}{\square}$", "$\sqrt{\square}$" and the superscript expressed implicitly. So the number of levels equals to 4. The values of L_i of each symbol are listed as follows.

- *Level* = 0: "5", "+", "$\frac{\square}{\square}$";
- *Level* = 1: "$\sqrt{\square}$", "*a*", "-", "*b*";
- *Level* = 2: "1", "-", "*c*";
- *Level* = 3: "2".

Table 1 shows the FDS data of the formula.

TABLE I. FDS DATA OF "$5 + \frac{\sqrt{1 - c^2}}{a - b}$"

C	*P*	*L*	*O*	*F*
5	1	0	0	0
+	2	0	1	0
\frac	3	0	1	0
\sqrt	4	1	1	1
1	5	2	0	6
-	6	2	1	6
c	7	2	0	6
2	8	3	0	2
a	9	1	0	5
-	10	1	1	5
b	11	1	0	5

From Table 1 we can find that the geometrical features of each symbol in the formula are marked clearly through the FDS data. We can employ them to find the index and retrieval features of formulae for realizing mathematical indexing and matching with high efficiency.

III. MATHEMATICAL INDEX MODEL BASED ON THE HIERARCHICAL FEATURES OF OPERATORS

With the help of FDS data of mathematical expressions, we can divide formulae into several classes and build their math index for realizing mathematical retrieval, which could effectively avoid excessive size of index resulted from large amounts of mathematical expressions.

A. Formula Feature Extraction

Through the analysis of mathematical expressions with FDS data, the hierarchical features are extracted for the classification of formulae through the following steps.

- Identification of two dimensional operators

  Assume $S_i(i=1,...,n)$ is a symbol in formula *ME* which contains *n* symbols. $A(S_i)=(C_i, P_i, L_i, O_i, F_i)$ is the FDS data of symbol S_i. $\{A(S_i)=(C_i, P_i, L_i, O_i, F_i)|_{i=1,...,n}\}$ is the FDS data of formula *ME*. Opt_{ME} is the set of the operators in formula *ME*. Optd_{ME} is the set of the two dimensional operators in formula *ME*. They are counted according to the following rule.

  If $O_i = 1$

  Then $S_i \in \mathrm{Opt}_{ME}$

  If $O_i = 1 \;\&\&\; F_{i+1} \neq 0$

  Then $S_i \in \mathrm{Optd}_{ME}$

- Calculation of the parameters of formulae

  Assume $\mathrm{Opt}_{MEj}(j=1,...,m)$ is the set of the Opt_{ME} in L_j in *ME* which has *m* levels.

  If $S_i \in \mathrm{Opt}_{ME} \;\&\&\; L_i = j$

  Then $S_i \in \mathrm{Opt}_{MEj}$

  The number of the operators in L_j is defined as $N_{\mathrm{Opt}j}$. The series number of S_i in Opt_{MEj} is defined as $\mathrm{Opt}_{MEj}(S_i)$.

  Assume $\mathrm{Optd}_{MEj}(j=1,...,m)$ is the set of the Optd_{ME} in L_j in *ME* which has *m* levels.

  If $S_i \in \mathrm{Optd}_{ME} \;\&\&\; L_i = j$

  Then $S_i \in \mathrm{Optd}_{MEj}$

  The number of the two dimensional operators is defined as $N_{\mathrm{Optd}j}$. The series number of S_i in Opt_{MEj} is defined as $\mathrm{Optd}_{MEj}(S_i)$.

- Construction of the hierarchical feature vectors of formulae

  Assume $\mathrm{Opt}_{MEj}=\{o_{tjp}(j=1,...,m;p=1,...,r)\}$ where *r* is the number of the operators in L_j of *ME*, $\mathrm{Optd}_{MEj}=\{o_{\mathrm{d}\,jl}(j=1,...,m;l=1,...,k)\}$ where *k* is the number of two dimensional operators in L_j of *ME*.

$$H_{ME}=[h_0,h_1,...,h_{t-1}]^{\mathrm{T}} \tag{2}$$

Where

$$h_0=\frac{\mathrm{Opt}_{ME0}(o_{\mathrm{d}01})+1}{N_{Opt0}+1} \tag{3}$$

$$h_1=\frac{\mathrm{Opt}_{ME0}(o_{\mathrm{d}0\,k/2})+1}{N_{Opt0}+1} \tag{4}$$

$$h_2=\frac{\mathrm{Opt}_{ME0}(o_{\mathrm{d}0\,k})+1}{N_{Opt0}+1} \tag{5}$$

$$h_3=\frac{\mathrm{Opt}_{ME0}(o_{\mathrm{dm}1})+1}{N_{Opt0}+1} \tag{6}$$

$$h_4=\frac{m}{N_{Opt0}+1} \tag{7}$$

… …

Through clustering analysis based on the hierarchical features of operators $H_{ME}=[h_0,h_1,...,h_{t-1}]^{\mathrm{T}}$, the mathematical expressions are divided into several classes called $T_0,T_1,...,T_{w-1}$ where *w* is the number of the classes.

B. Mathematical Index Construction

The math index based on hierarchical features of operators contains three layers: hierarchical classes, index and inverted index. Its structure is shown in Fig. 3.

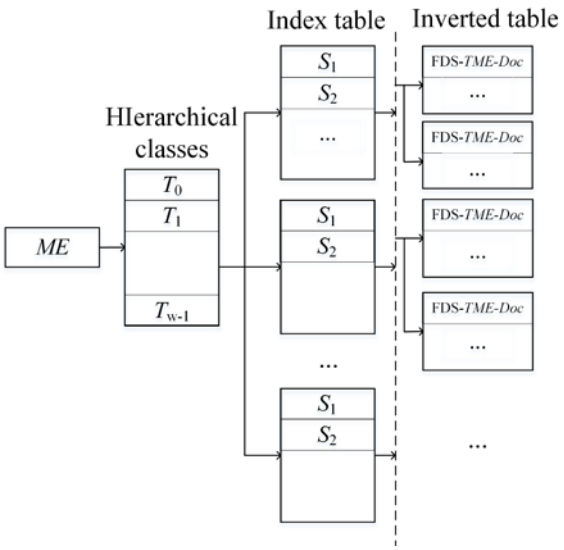


Fig. 3. Structure of the mathematical index based on hierarchical features of operators of formulae

Compared with traditional full text retrieval model, this index adds the hierarchical feature classes before the symbol match operations. Through the classifying operation to the query expression, the class it should belong could be found and the matching operation will be carried out in the corresponding inverted index firstly. This is helpful for the efficiency of searching operation.

IV. Experimental Result and Analysis

The mathematical index proposed in the paper was constructed and the corresponding match experiments were carried out. There are totally 134199 formulae in our mathematical expression library including almost all types of formulae such as ordinary one dimensional expressions, fractions, radicals, integrals, differentials, superscripts and subscripts, and so on.

With the analysis method we proposed, the level feature data of formulae are obtained as shown in Fig. 4.

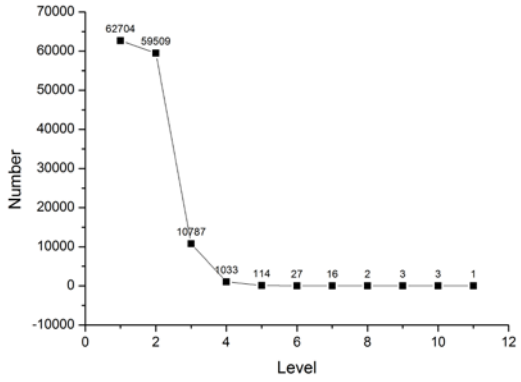


Fig. 4. Level feature data of formulae

The hierarchical feature data based on the operators, especially two dimensional operators, was also extracted as shown in Fig. 5.

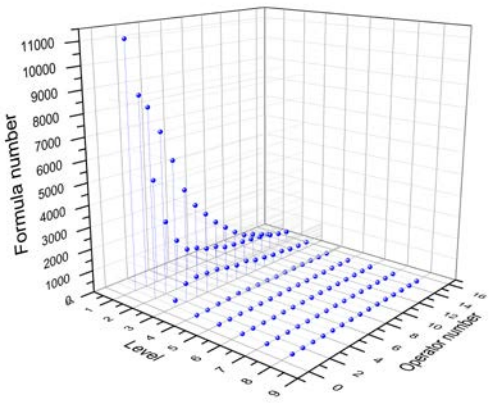


Fig. 5. The hierarchical feature data of formulae based on the operators

The size of the three layer index is a little bigger than the two layers mathematical index, which is at the MB level.

V. Conclusions

In this paper, a mathematical expression index is designed for improving the efficiency of math retrieval. It employed the hierarchical features of formula's operators to divided the formulae into several classes for establish index respectively. It still has some shortages to be improved. For example, it is more adapt to the exactly matching of mathematical expressions because the hierarchical features of formulae are over all a global feature, which might result in the loss of local information of expressions. So our further work is to build the connect channels among formula classes so as to improve the matching ability of the math index.

Acknowledgment

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

References

[1] B R. Miller, "Three years of DLMF: web, math and search,". International Conference on Intelligent Computer Mathematics. Springer Berlin Heidelberg, 2013, pp. 288-295.

[2] R. Miner, R. Munavalli, "An Approach to mathematical search through query formulation and data normalization," Lecture Notes in Computer Science: Towards Mechanized Mathematical Assistants. Springer Berlin Heidelberg, 2007, vol. 4573: pp. 342-355.

[3] P. Libbrecht and E. Melis. "Methods to access and retrieve mathematical content in ActiveMath," Mathematical Software - ICMS 2006, Lecture Notes in Computer Science, Vol. 4151, 2006, pp. 331-342.

[4] J. Mišutka, L. Galamboš, "System Description: EgoMath2 As a tool for mathematical searching on Wikipedia.org," Intelligent Computer Mathematics, Lecture Notes in Computer Science, 2011, Vol. 6824, pp. 307-309.

[5] R. Hambasan, M. Kohlhase, C. C. Prodescu, "MathWebSearch at NTCIR-11," NTCIR. 2014.

[6] X. Hu, L. C. Gao, X. Y. Lin, Z. Tang, X. F. Lin, J. B. Baker, "WikiMirs: a mathematical information retrieval system for Wikipedia," In: Proceedings of the 13th ACM/IEEE-CS joint conference on digital libraries. ACM, 2013, pp. 11-20.

[7] X. Y. Lin, L. C. Gao, X. Hu, , Z. Tang, Y. N. Xiao, X. Z. Liu, "A mathematics retrieval system for formulae in layout presentations," Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval (SIGIR '14). ACM, 2014, pp. 697-706.

[8] http://www.latexsearch.com/, 2017.7.1.

[9] S. Hong, W. Su, H. Lin, "Functional classification study for mathematical formulas retrieval," 2016 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD). IEEE, 2016, pp. 99-104.

[10] T. Schellenberg, B. Yuan, R. Zanibbi, "Layout-based substitution tree indexing and retrieval for mathematical expressions," //IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, 2012, 8297:pp. 829701-1-829701-8.

[11] R. Zanibbi, B. Yuan, "Keyword and image-based retrieval of mathematical expressions," //IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, 2011, pp. 78740I-78740I-9.

[12] X. D. Tian, S. Q. Yang, X. F. Li, "An indexing method of mathematical expression retrieval", Proceedings of the 2013 3nd International Conference on Computer Science and Network Technology, pp 574-578.