Research on Mathematical Formula Knowledge Base for Formula Recognition

Zhijun Guo
Teaching Affairs Department
The Open University of China
No. 75, Fuxing Road, Beijing 100039 P. R. China
guozhijun@ouchn.edu.cn

Yao Liu
Engineering Research Center
Institute of Scientific and Technical Information of China
No. 15, Fuxing Road, Beijing 100038 P. R. China
liuy@istic.ac.cn

Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University)
Fuzhou 350121, P. R. China

***Abstract*—A mathematical formula image must be able to reproduce its layout, syntax structure, and semantic meaning before it can be widely used. In order to achieve an accurate reproduction of mathematical formulas, the first important thing is to establish an accurate and comprehensive formula model. Based on the ontology thought, this paper presents a mathematical formula model and builds a mathematical knowledge base based on this model. The mathematical formula knowledge base uses an ontology to store mathematical formulas, so that the mathematical symbols, formula structures, and formula semantics are effectively combined to form a semantic network. The mathematical formula knowledge base can be used for the recognition of mathematical formulas. Furthermore, it can accurately describe the computational semantics expressed by the mathematical formula.**

Keywords—Mathematical formula, Ontology, Knowledge base, Formula recognition

I. INTRODUCTION

The study of mathematical formulas recognition originated from Anderson's paper in 1967, which, for the first time, proposed the problem of mathematical formula recognition [1]. However, the mathematical formula recognition research progress has been relatively slow after then. As OCR technology has matured, the research of formula recognition has gradually increased, and great progress has been made. The current researches are mainly focused on the recognition and reproduction, and the researches on the mathematical models and the storage methods are relatively less [2,3,4]. However, in order to apply the identified mathematical formula, it needs to effectively store and retrieve formulas. For the sake of effectively storing and retrieving formulas, it is necessary to construct a reasonable mathematical formula knowledge base based on accurately describing the layouts, syntactic structures, and semantic meaning.

Ontology is a conceptual system that describes the domain knowledge in specified fields. It identifies the basic concepts in the field and provides a common understanding of the knowledge in the field. Furthermore, it gives a clear definition of the interrelationships between these concepts at diversity levels. It is therefore generally assumed that the ontology is a description of the concepts and their relationship of a specified domain. It can effectively combine mathematical symbols, formula structures, and formula semantics to form a semantic network using the ontology to describe the mathematical formulas. In this study, the ontology thought is introduced into the mathematical formula model, so that the mathematical formula model can describe the semantic meaning of mathematical formula more accurately, so as to promote the recognition and application for the mathematical formula. The organizational structure of this paper is as follows: Firstly, the mathematical formula model based on ontology is described, and then the method of constructing mathematical formula knowledge base is put forward. Finally, the platform for constructing mathematical formula knowledge base is introduced.

II. MATHEMATICAL FORMULA MODEL

To use the ontology to describe the mathematical formula, it firstly needs to determine the basic concepts involved in the mathematical formula and then find out the relationship between the concepts. That is, between the concepts through which the attributes are associated. The mathematical formula model proposed in this paper mainly involves the following concepts: operand, operator, basic structure, composite structure, and mathematical formula [5-11].

A. Operand

Operands are composed of algebraic symbols such as numbers, letters, and Greek letters. They represent numbers, variables, and so on in mathematical formulas. The properties of the operands in the mathematical formula model are as shown in table 1. There are four types of operands in this model: numbers, letters, Greek letters, and punctuation marks.

978-1-5386-7325-6/18/$31.00 ©2018 IEEE
DOI 10.1109/WI.2018.00-27

TABLE 1. ATTRIBUTES OF THE OPERAND

Attributes	ID
	Name
	Type
	Symbol

B. Operator

Operators contain signs of operation, function names, and some special symbols. In a mathematical formula, the operators represent an operational relationship to one or more operands or a particular mathematical rule. The properties of the operators in the mathematical formula model are as shown in table 2.

TABLE 2. ATTRIBUTES OF THE OPERATOR

Attributes	ID
	Name
	Type
	Category
	Priority
	Structure
	Symbol
	Semantic

The type of operator is to classify the operators by type of operation. It includes a description, a definition, a combination, a relationship, an arithmetic, and a function. A category refers to the number of operands that can be combined with a single operator to express a mathematical relationship or a particular mathematical rule. The range of category includes single, double, single or double, and triple or more.

The priority refers to the priority of the operator at the same level within the same expression. High priority means early operation. For an expression that contains only one operator, it does not need to consider the priority of the operator. The following table 3 lists the priorities between the operators.

TABLE 3. PRIORITY OF OPERATORS

	Relationship	**Definition**	**Combination**	**Arithmetic**
Relationship		>	>	>
Definition	<		>	>
Combination	<	<		>
Arithmetic	<	<	<	

The structure defines the location of the subexpression that can be combined with the operator. The location refers to up, down, left, right, top left, bottom left, top right, bottom right, and inside. The structure can be represented by one or more subexpression locations.

C. Basic Structure

The basic structure is an expression that contains only one operand. Normally, a basic structure contains only one operator and does not contain operators in some special cases, such as power operations. An expression is a logical element that makes up a mathematical formula. A mathematical formula can be a single expression or combination of multiple expressions. The forms of expressions are as follows: 1. A single operand or a combination of operands; 2. A combination of a single operator and its operands; 3. A combination of joins by 1 and/or 2. The attributes of the basic structure are as shown in table 4.

TABLE 4. ATTRIBUTES OF THE BASIC STRUCTURE

Attributes	ID
	Name
	Type
	Operator
	Layout description
	Semantic

There are two types of basic structure: class A and class B [6]. Class A indicates that the layout position of the subexpressions is entirely determined by the mathematical notation, such as expressions consisting of a single character, fractional expressions, and root expressions. Class B indicates that the layout position of the subexpressions is determined by all symbols, such as function expressions, integral expressions, and so on.

The Layout description defines the location of the subexpressions that can be combined with the operator. There are nine subexpression positions which are up, down, left, right, top left, bottom left, top right, bottom right, and inside. The semantic refers to the computational meaning of the basic structure.

D. Composite Structure

A composite structure consists of more than one basic structure. It contains at least two operators, and it is a combination of multiple basic structures. The attributes of the composite structure are as shown in table 5.

TABLE 5. ATTRIBUTES OF THE COMPOSITE STRUCTURE

Attributes	ID
	Name
	Operators
	Basic Structures
	Layout description
	Semantic

E. Mathematical Formula

Mathematical formulas consist of numbers, algebraic symbols, and operator symbols. They represent quantity, variables, operations and mathematical rules [7]. The mathematical formulas referred to in this paper are composed of basic structures or composite structures and operands. The attributes of the mathematical formula are as shown in table 6.

TABLE 6. ATTRIBUTES OF THE MATHEMATICAL FORMULA

Attributes	ID
	Name
	Operators

	Operands
	Layout description
	Semantic

III. CONSTRUCTION OF KNOWLEDGE BASE

The construction of mathematical formula knowledge base is based on the above mathematical formula model to build mathematical formula ontology. The ontology construction method includes manual construction and automatic construction. Manual construction is characterized by precision while automatic construction is characterized by fast. In order to take full advantage of each build method, this paper constructs the mathematical formula knowledge base using mixed ways. That is, using the manual method as well as the automatic method. The construction of the operands, operators and basic structures is done using the manual method. The construction of the composite structures is done using the automatic method. To show a mathematical formula in a document, it needs to describe the layout of the mathematical formula. That is to describe the external structure of the formula so that people can understand its mathematical meaning. Operators and expressions are more complex and diverse than numbers and letters. Latex is a TEX-based typography system that supports a wide variety of mathematical symbols. It can generate complex mathematical symbols and formulas. This article uses Latex to describe Greek letters, operators, and expressions.

A. Construction of operand and operator

Operands include numbers, capital English letters, lowercase English letters, capital Greek letters, and lowercase Greek letters in 100 characters. The numbers and the letters are stored in ASCII, and the Greek letters are stored with Latex.

Mathematical operators contain a variety of mathematical symbols, such as a plus sign, minus sign, times sign, division sign, integral sign, and radical sign etc. This study uses the Latex to the describe operator. The operators in the mathematical formula knowledge base contain all the mathematical symbols that Latex supports. A sample of the operators and their Latex description are as shown in table 7.

TABLE 7. A SAMPLE OF THE OPERATORS AND THEIR LATEX DESCRIPTION

Operator	Latex	Operator	Latex	Operator	Latex
∑	\sum	≡	\equiv	∈	\in
∏	\prod	≠	\neq	∉	\notin

In order to describe the computational meaning expressed by the mathematical formula, the semantic information expressed by the operator must be given. MathML is used to describe the structure and content of mathematical formulas. It provides two types of markers that describe mathematical formulas: Presentation tags and Content tags. The Presentation tags describe the two-dimensional layout information of the mathematical formulas, focusing on the appearance. The Content tags describe the computational meaning of the mathematical formulas, concentrating on the meaning of the expression. In this paper, it uses MathML's Content tags to describe the semantic information of mathematical formulas. The sample of content tags are as shown in table 8.

TABLE 8. SAMPLE OF CONTENT TAGS

Tag	Explanation	Tag	Explanation
<plus>	plus	<divide>	divide
<subtract>	subtract	<multiply>	multiply

B. Construction of basic structure

The basic structure of the mathematical formula is used to describe the spatial relationship between the operator and its subexpression. The operator acts as the core symbol of a basic structure and is responsible for the operation between the operands. For each operator, the position of its subexpression can be obtained by statistical analysis and prior knowledge. For example, the position of its subexpression for the symbol "+" can only be on its left and right, and not on the top or bottom. This study summarizes the possible positions of the subexpressions of commonly used operators and builds the basic structure knowledge base based on them. The example is shown in figure 1.

Fig. 1. Common basic structure example

C. Construction of composite structure

Each mathematical formula that contains two or more operators corresponds to a composite structure. The basic structures of mathematical formulas are limited in quantity and can be obtained manually by statistical analysis and prior knowledge. But the composite structures of mathematical formulas cannot be manually and effectively accessed because of its complex shape and mass of the number. The composite structure and the mathematical formula have a one-to-many relationship, so the composite structure can be obtained by analyzing the formula structure in the process of formula recognition. The main steps include the following: 1. obtaining the characters of the formula image and its two-dimensional structure information; 2. analyzing the structure of the formula based on the mathematical formula knowledge base; 3. acquiring the new composite structure based on the mathematical formula knowledge base [9]. The process of acquiring the composite structure automatically is shown in figure 2.

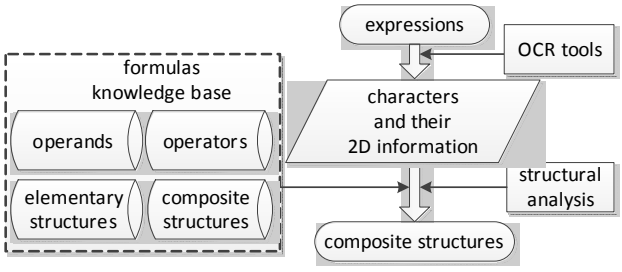

Fig. 2. Composite structure acquisition workflow

IV. KNOWLEDGE BASE CONSTRUCTION PLATFORM

Based on the above studies, combined with the most current information technology, this paper constructed an online formula knowledge base construction platform designed to build the mathematical formula knowledge base, and verify the mathematical formula model proposed by this paper.

A. Overall framework

The mathematical formula knowledge base construction platform mainly implements the editing of the related attributes of operators and operands, as well as the maintenance of the basic structures and the automatic extraction of composite structures. The platform also provides user management functions to manage users. The platform based on the Web, so that the users can use it through the internet. The overall framework of the platform is as shown in figure 3.

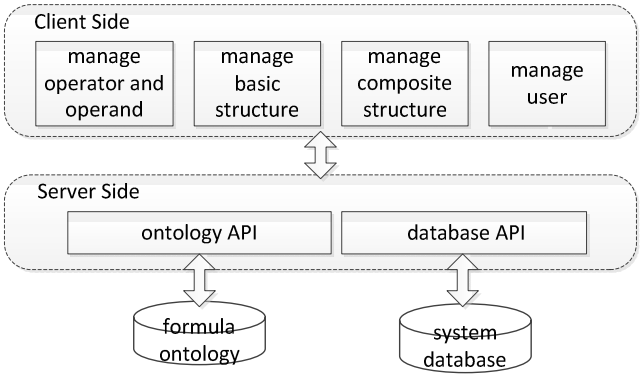

Fig. 3. Overall framework

B. Main functions

In addition to the automatic extraction of complex structures, the platform also needs to maintain the attributes of the operator and operand used in the extraction process. According to the above requirements, the platform provides the following main functions: maintenance of operators and operands, maintenance of the basic structure, automatic extraction of complex structures, and management of users.

V. CONCLUSIONS

In order to accurately reproduce the layout, syntactic structure, and semantic meaning of mathematical formulas, this paper builds a mathematical formula model based on ontology. The model makes a detailed classification of frequently-used symbols of the mathematical formula and describes the core concepts used by the model. The model uses properties of concepts to effectively connect core concepts. Based on the model, a mathematical knowledge base is constructed. The mathematical knowledge base constructed in this study combines mathematical symbols, formula structures, and formula semantics to form a semantic network. The knowledge base can be used for the recognition of mathematical formulas, and it can accurately describe the computational semantics expressed by the mathematical formula. In this paper, the third-party tool is used as the extraction tool of the mathematical formula in the process of constructing the formula knowledge base. The efficiency and accuracy of the composite structure are still to be improved due to the limitation of the tool and the complexity of the two-dimensional structure of the mathematical formula. The future study of this work is to improve the extraction efficiency and accuracy of the composite structure and enrich the formula knowledge base.

ACKNOWLEDGMENTS

This work is supported by the Open Fund Project of Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) (No. MJUKF201739). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Anderson, R. H. (1967, August). Syntax-directed recognition of hand-printed two-dimensional mathematics. In Symposium on Interactive Systems for Experimental Applied Mathematics: Proceedings of the Association for Computing Machinery Inc. Symposium (pp. 436-459). ACM.

[2] Jiao Chen (2005). Structure Description and Analysis of Mathematical Expressions[D]. Nankai University.

[3] Na Zhao (2007). Structure Analysis, Understanding and Reproduction of Mathematical Expressions Images[D]. Nankai University.

[4] Guangshun Shi, Cui Xiao, Qingren Wang (2008). Structure Understanding and Reproduction of Mathematical Expressions Images [J]. CAAI Transactions on Intelligent Systems, 3(5):401-407.

[5] Cui Xiao (2009). Research on Systematic Design and Key Methods of Mathematical Formula Structure Analysis [D]. Nankai University.

[6] Ashida K, Okamoto M, Imai H, et al (2006). Performance Evaluation of a Mathematical Formula Recognition System with a large scale of printed formula images[C]// International Conference on Document Image Analysis for Libraries. IEEE Computer Society, 320-331.

[7] Jianming Jin, Hongying Jiang (2001). Research Status of Typeset Mathematical Expressions Processing[C]// China Intelligent Automation Conference.

[8] Zhijun Guo, Yao Liu (2015). The Key Technology Research of Mathematical Formula Recognition and Construction of Application Platform[J]. ICIC Express Letters Part B: Applications. 697-701.

[9] Ruijia Wang, Yao Liu (2012). Semantic analysis of multi-modal features in scientific and technical literature[J]. ICIC Express Letters Part B: Applications. 901-908.

[10] Simistira F, Katsouros V, Carayannis G (2015). Recognition of online handwritten mathematical formulas using probabilistic SVMs and stochastic context free grammars[J]. Pattern Recognition Letters, 53:85-92.

[11] Julca-Aguilar F, Mouchère H, Viard-Gaudin C, et al (2015). Top-Down Online Handwritten Mathematical Expression Parsing with Graph Grammar[M]// Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Springer International Publishing.