On Ambiguity Issues of Converting LaTeX Mathematical Formula to Content MathML

Kai Wang, Xinfu Li(✉), and Xuedong Tian

Hebei Key Laboratory of Machine Learning and Computational Intelligence, College of Computer Science and Technology, Hebei University, Shijiazhuang, China
mc_lxf@l26.com

Abstract. Facing the demand of providing retrieval result with rich semantic information for users in math searching, mathematical formulas of LaTeX are usually converted to Content MathML. For the problem of ambiguity formulas in the process of conversion, a method of semantic disambiguation for mathematics formulas which is based on the operator context was proposed. At first, ambiguity operator was found according to the ambiguity operator mapping table. Then, the ambiguity operator context is got through the array traversal. At last, the specific meaning was conjectured according to the ambiguity operator context. The experimental results show that compared with the type-system the method can make up for its disadvantages in simple formulas aspect and gets a higher average accuracy. In practical application, this method can effectively solve the problem of ambiguity formulas in the process of conversion.

Keywords: Mathematic formula conversion · LaTeX · Content MathML · Ambiguity formula · Semantic of formula

1 Introduction

Mathematical formula is more and more widely used on the Web. As the basis of mathematical formula retrieval system, format conversion is particularly important. The forms of LaTeX and MathML (Mathematical Markup Language) [1] formulas have been rapidly developed with their unique characteristics. Therefore, the issue of converting LaTeX mathematics formulas to Content MathML becomes a top priority in related fields. It would be an enormous job to systematically codify most of mathematics by hand – a task that can never be complete. The key problem is how to solve the problem of ambiguity formula.

Many scholars had done related research for mathematical formula tree structure, format conversion and ambiguity formula. Guan [2] extracted the tree structure of MathML formula and realized the related operations. Nie et al. [3] put forward an algorithm based on index and had carried on the concrete analysis about the tree structure of LaTeX formula. Zhang [4] used the principle of list and stack and combined with the operator priority realized the conversion of Infix to Content MathML. Zhao [5] proposed a method with higher recall ratio and precision ratio about the standardization of mathematical formula and semantic retrieval. However, there is no

S. Guo et al. (Eds.): CollaborateCom 2015, LNICST 163, pp. 289–295, 2016.
DOI: 10.1007/978-3-319-28910-6_26

study for the problem of ambiguity formula. Ting Zhang et al. [6] invented a formula translator called MathEdit. They fulfilled the conversion among Presentation MathML, Content MathML and Infix, which provided a great convenience for different forms of mathematical formula input, output, and storage. Su [7] came up with a kind of method based on the binary representation of the complexity of computation and realized the conversion among several common mathematical formula formats. The literature [8, 9] respectively introduced a method of conversion, that is the Presentation MathML to LaTeX and LaTeX to Presentation MathML. Doush et al. [10] put forward a framework of adding semantic information in the mathematical expression and realized the conversion of Presentation MathML to Content MathML using a method of RDFa (Resource Description Framework in attributes). Cai et al. [11] studied the problem of ambiguity mathematical formula transformation about Presentation MathML to Content MathML. The problem of LaTeX mathematical formula to the Content MathML ambiguity transformation was not involved.

According to the inspiration of scholars both at home and abroad, an ambiguity conjecture method for mathematical formulas based on operator context was proposed to deal with the ambiguity problems in the process of conversion about LaTeX mathematical formula to Content MathML.

2 Ambiguity Mathematical Formula

Some mathematical symbols of LaTeX have several meanings although they have unique representation. We need to clarify their specific meaning in the process of conversion. This kind of formulas with ambiguity operators are called ambiguity mathematical formula. LaTeX and Content MathML mapping table is listed in Table 1.

Table 1. LaTeX and content MathML mapping table

Ambiguity operator	LaTeX	Meaning	Content MathML
×	\times	product	<times/>
		vector product	<vectorproduct/>
		cartesian product	<cartesianproduct/>
superscript(T)	^T	power	<power/> … < ci > T</ci>
		transposition	<transpose/> … < ci > T</ci>
superscript(-1)	^{-1}	power	<power/> … < minus/> <cn > -1 </cn>
		inverse function inverse of a matrix	<inverse/>
| |	\[\left|{ } \right|\]	absolute	<abs/>
		aggregative card	<card/>
{ }	\[\left\{{ } \right\}\]	set	<set>
		grouping	<apply>

(*Continued*)

Table 1. (*Continued*)

Ambiguity operator	LaTeX	Meaning	Content MathML
[ ]	\[\left[{} \right]\]	closed interval	<interval>
		grouping	<apply>
( )	\[\left({} \right)\]	open interval	<interval>
		list	<list>
		grouping	<apply>
		vector	<vector>
d*	d*	differential coefficient	<diff/>
		integral	<int/> <bvar>

3 Ambiguity Conjecture Method

This paper proposes an ambiguity conjecture method for mathematical formula based on operator context. When finding ambiguity operator extracts its context first, then according to the different forms of contexts to conjecture the specific meaning of the operator in the process of conversion.

Using MathType to get LaTeX mathematical formula first, then tokenizing the formula into the smallest item and saving them to array. If it is an ambiguity operator it's position and context could be retrieved from the array. There are kinds of contexts. Some common forms are shown in Table 2.

Table 2. Some common forms of the context of ambiguity operator

Order	Context forms	Examples	Specific examples	Meaning
1	uppercase	*A*,*B* et al.	$A \times B$	product or vector product or cartesian product
2	lowercase	*a*, *b* et al.	$a \times b$	product
3	number	1,123 et al.	123×456	product
4	expression	$x + y$, $a*b$ et al.	$(x + y) \times z$	product
5	Greek letter	α, βet al	$\alpha \times \beta$	product or vector product
6	special form	Null, "T" et al.	A^T	power or transposition
7	function	sin *x*, f (*x*) et al.	$\sin x \times \cos x$	product

The algorithm based on operator context is designed according to the same ambiguity operator has different meanings with different context and combines the ambiguity mathematical properties of the operator itself. A program flow chart is shown in Fig. 1.

Using pseudo code to describe the context algorithm:

```
IF the operator is "\times" THEN
          DO getting "above" and "below";
IF above is "uppercase" and below is "uppercase" THEN
           DO related translation;
ELSE IF above is "number" and below is "lowercase" THEN
           DO related translation;
ELSE IF the operator is "^" THEN
           DO getting "above" and "below";
IF above is "function" and below is "T" THEN
           DO related translation;
ELSE IF above is "expression" and below is "number"
THEN
           DO related translation;
END
```

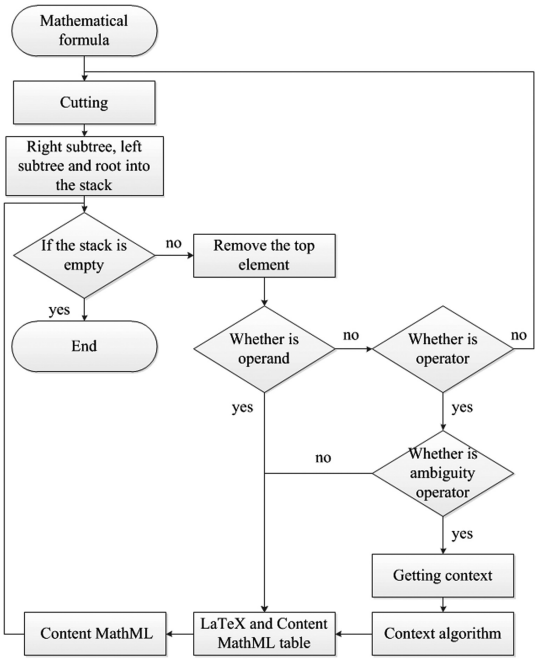

Fig. 1. Program flow chart

4 Experimental Results and Analysis

In order to verify the validity of the method, this paper gets the ambiguity mathematical formulas from the public data set [12]. There are 483,874 mathematical formulas in total. This paper using the simple random sampling extracts 1,000 ambiguity formulas to carry out the experiment and artificially contrast. The experimental result is shown in Table 3.

Table 3. Results of the experiment

Type	Meaning	Total formulas	Correct formulas	Accuracy ratio (%)	Simple formula	Correct formulas	Accuracy ratio (%)
×	product	723	660	91.3	216	149	69.0
	matrix product	231	183	79.2			
	cartesian product	46	35	76.1			
^{T}	power	463	423	91.4	240	170	70.8
	transform	537	483	89.9			
^{-1}	power	189	176	93.1	35	28	80.0
	inverse matrix	48	45	93.8			
	inverse function	30	24	80.0			
| |	absolute	903	815	90.3	178	129	72.5
	card	97	76	78.4			
{ }	set	135	112	83.0	–	–	–
	grouping	865	787	91.0			
[ ]	closed interval	117	110	94.0	–	–	–
	grouping	883	770	87.2			
( )	open interval	106	96	90.6	–	–	–
	grouping	782	702	89.8			
	list	23	11	47.8			
	vector	89	55	61.8			
d*	integral	763	721	94.5	–	–	–
	differential coefficient	237	220	92.8			

In the experiment, there are 7,267 ambiguity mathematical formulas of LaTeX in total. The average conversion accuracy is 89.5 %. However, the simple formulas of accuracy ratio are far below the value.

Through the analysis of the results is not hard to find that the formulas of incorrectly conjecturing most are simple ambiguity formulas which led to a low accuracy of simple ambiguity formulas. Simple ambiguity formulas have in common is that their context is too single, so it is different to judge the true meaning of ambiguity operator. If natural language context of formula can be combined, the ambiguity mathematical formula accuracy ratio can be improved in the actual process of conversion.

5 Conclusion

This paper analyzed a problem about ambiguity formulas in the process of LaTeX to Content MathML conversion. The experimental results show that compared with the type-system the method makes up for its disadvantages in dealing with simple formulas and has a higher average accuracy. In practical application, this method can effectively solve the problem of ambiguity formulas in the process of conversion.

There is no denying that the deficiencies are in the system due to insufficient time. The method is simple and needs to do a lot of test. The next step will focus on how to combine natural language context to further improve the ambiguity mathematical formula transformational accuracy ratio. LaTeX mathematics formula to Content MathML is still in the immature stage and needs to do more researches.

Acknowledgments. This work is supported by National Natural Science Foundation of China (No. 61375075), Natural Science Foundation of Hebei Province (No. F2013201134).

References

1. W3C Math Working Group. Mathematical Markup Language (MathML) version 3.0 2nd edn. W3C Recommendation, 10 April 2014. http://www.w3.org/TR/MathML/
2. Guan, M.: The research of mathematical formula indexing based on MathML. Hebei University, Hebei (2013)
3. Nie, J., Chen, T., Fu, H.: Research and implementation of internet mathematic formula search engine based on Latex. J. Comput. Appl. **30**(2), 312–313 (2010)
4. Zhang, T.: Research and Implementation of Web-based mathematical expressions translation. Lanzhou University, Lanzhou (2009)
5. Zhao, L.: Research on the method and technology of mathematical formulas semantic retrieving based on ontology. Nankai University, Tianjin (2011)
6. Zhang, T., Li, L., Su, W., et al.: A mathematical formula converter based on mathedit. Comput. Appl. Softw. **27**(1), 14–16 (2010)
7. Su, W.: Research on web-based input and accessibility of mathematical expressions. Lanzhou University, Lanzhou (2010)
8. Stamerjohanns, H., Ginev, D., David, C., Misev, D., Zamdzhiev, V., Kohlhase, M.: Mathml-aware article conversion from LaTeX (English). In: Petr, S. (ed.) Towards a Digital Mathematics Library, Grand Bend, 8–9 July 2009, pp. 109–120. Masaryk University Press, Brno (2009)
9. Woodall, D.R.: LaTeX MathML: translating LaTeX math notation dynamically to presentation MathML (2010)

10. Doush, L.A., Alkhateeb, F., Maghayreh, E.A.: Towards meaningful mathematical expressions in e-learning. In: Proceedings of the 1st International Conference on Intelligent Semantic Web-Services and Applications, Amman (2010)
11. Cai, C., Su, W., Li, L.: On key issues of converting presentation mathematics formulas to content. Comput. Appl. Softw. **29**(8), 30–33 (2012)
12. NICIR 11 Math Wikipedia Task. http://ntcir11-wmc.nii.ac.jp/index.php/NTCIR-11-Math-Wikipedia-Task