A Novel Similarity-Search method for Mathematical Content in *LaTeX* Markup

Wei Zhong, Hui Fang
Dept. of Electrical and Computer Engineering
University of Delaware
Newark, DE USA
{zhongwei, hfang}@udel.edu

ABSTRACT

A relaxed structural matching search method, along with a symbolic similarity measurement algorithm for mathematical content search is proposed. Our approach uses an intermediate tree representation to capture structural information of mathematical expression, and based on a previous idea which indexes math expression structure through tree leaf-root paths, we further describe an advanced AND search method in a formal way. This search method can be used to test query/document subexpression isomorphism or evaluate the symbolic similarity between math expressions with consideration of their α-equivalence. For the purpose of evaluation, we also implement a search engine based on our idea.

Categories and Subject Descriptors

H.3 [**Information Search and Retrieval**]: Miscellaneous

General Terms

Algorithms

Keywords

mathematical searching, language processing, search engine

1. INTRODUCTION

With *MathJax* becoming popular, more and more LaTeX markups can be crawled directly from many websites. In order to search those mathematical language in LaTeX markups, a search method that can respect the properties of math expression needs to be developed. Although many researches have been conducted to retrieve information in structured content (e.g. *MathML*), information retrieval on LaTeX math content is still not well-studied or exhaustively covered by mainstream IR research, compared to that on general text.

Unlike general text content, mathematical language, by its nature, has many differences from other textual documents, there are a number of new problems in measuring mathematical expression similarity. Among those problems, we know one math expression can be transformed to alternative forms, e.g. $\frac{a+b}{c}$ and $\frac{a}{c}+\frac{b}{c}$ should be considered as semantically identical. To identify those variations requires search engine to apply mathematical transformation rules to a query in order to obtain all forms of relevant expressions. Further, math expressions with the same evaluated value may also be considered relevant, in this case, $\sin(\frac{\pi}{2})$ and 1 are equivalent. Some computational search engines (e.g. *Symbolab* and *WolframAlpha*) are aware of these problems. But sometimes we need to rely on conventions and context to distinguish expressions such as $f(a+b)$ and $c(a+b)$, because the symbol f in the former expression is likely to represent function instead of a variable, in addition, expression such as f^{-1} can either be reciprocal or an inverse function. Moreover, a higher level of understanding of mathematic knowledge may be required for math-aware search engine to find the results for queries such as "find an article related to the four color theorem" (from NTCIR-10 Math topics [1]).

Yet the problems addressed above are not considered in this paper, instead, we are focusing on the aspects which does not require a "good understanding" of mathematics, we target our research domain to be the following: The first is structural similarity. For example, $ax+(b+c)$ is not equivalent to $(a+b)x+c$ although they have the same set of symbols, this is because their structural difference. However, as the position of operands in math expression can be commutative in some cases, structural similarity is often measured by substructure isomorphism if we use operation tree [2] to represent math expressions. The second is symbolic similarity, with the consideration of α-equivalence. We know that symbols can be used interchangeably in each math formula to express the same meaning, e.g. $a^2+b^2=c^2$ and $x^2+y^2=z^2$. Nevertheless, we still weight symbolic similarity sometimes, for instance, $E=mc^2$ is considered more meaningful when exact symbols are used rather than just being structurally identical with $y=ax^2$. On the other hand, we should also weight α-equivalent expressions more, that is, changes of symbols in expression preserve more syntactic similarity when changes are made by substitution, e.g. for query $x(1+x)$, expression $a(1+a)$ are considered more relevant than $a(1+b)$. Because the "bond variable" x and a here are at the same positions and both supposed to represent the same value. All the points addressed here makes transitional IR methods (e.g. bag of words model and tf-idf weighting) deficient to handle math content.

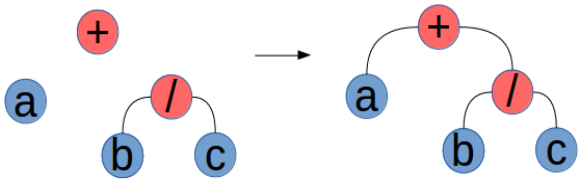

Figure 1: Example of Sub-tree generation for the addition grammar.

2. RELATED WORK

Similarity/boolean search for mathematical content is not a new topic, conference in this topic is getting increasingly research attention and the proposed systems have progressed considerably [3]. DLMF project from NIST [4, 5] and MIaS system [6, 7, 8], notably, use text-based approaches and utilize existing models to deal with math content on top of existing IR tools (such as *Apache Lucene*). They are commonly using augmentation and normalization (by ordering the subexpressions) to enumerate and represent all possible sequences of commutative operands. However, augmentation usually requires to index combination of both symbols (e.g. a and b) and unified items (*id*, *const*) in different levels of math expression, thus implies inefficient storage space usage for complex expressions. MWS [9, 10, 11] takes a *automatic theorem proving* approach and uses *term indexing* [12] to minimize the cost of unification algorithm which is able to find if two expressions are equivalent. However, their index relies on RAM memory and consumes a considerable space [11], and it indexes all sub-terms of a formula [9].

3. REFERENCES

[1] Topics for the ntcir-10 math task full-text search queries. http://ntcir-math.nii.ac.jp/wp-content/blogs.dir/13/files/2014/02/NTCIR10-math-topics.pdf. Accessed: 2015-03-31.

[2] Richard Zanibbi and Dorothea Blostein. Recognition and retrieval of mathematical expressions. *International Journal on Document Analysis and Recognition (IJDAR)*, 15(4):331–357, 2012.

[3] Akiko Aizawa, Michael Kohlhase, and Iadh Ounis. Ntcir-11 math-2 task overview. *The 11th NTCIR Conference*, 2014.

[4] Miller B. and Youssef A. Technical aspects of the digital library of mathematical functions. *Annals of Mathematics and Artificial Intelligence 38(1-3), 121âĂŞ136*, 2003.

[5] Youssef A. Information search and retrieval of mathematical contents: Issues and methods. *The ISCA 14th IntâĂŹl Conf. on Intelligent and Adaptive Systems and Software Engineering (IASSE 2005)*, 2005.

[6] Petr Sojka and Martin L. Indexing and searching mathematics in digital libraries. *Intelligent Computer Mathematics*, 6824:228–243, 2011.

[7] Petr Sojka and Martin L. The art of mathematics retrieval. *ACM Conference on Document Engineering, DocEng 2011*, 2011.

[8] Martin L. Evaluation of mathematics retrieval. Master's thesis, Masarykova University, 2013.

[9] Michael Kohlhase and Ioan A. SÂÿucan. A search engine for mathematical formulae. In *Proc. of Artificial Intelligence and Symbolic Computation, number 4120 in LNAI*, pages 241–253. Springer, 2006.

[10] Michael Kohlhase. Mathwebsearch 0.4 a semantic search engine for mathematics.

[11] Michael Kohlhase. Mathwebsearch 0.5: Scaling an open formula search engine.

[12] Peter Graf. *Term Indexing*. Springer Verlag, 1996.