AN EFFICIENT SIMILARITY BASED SEARCH ENGINE FOR MATHEMATICAL CONTENT IN LaTeX MARKUP

by

Wei Zhong

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Electrical and Computer Engineering

Spring 2015

AN EFFICIENT SIMILARITY BASED SEARCH ENGINE FOR MATHEMATICAL CONTENT IN LaTeX MARKUP

by

Wei Zhong

Approved: _____

Fouad E. Kiamilev, Ph.D.
Professor in charge of thesis on behalf of the Advisory Committee

Approved: _____

Kenneth E. Barner, Ph.D.
Chair of the Department of Electrical and Computer Engineering

Approved: _____

Babatunde A. Ogunnaike, Ph.D.
Dean of the College of Engineering

Approved: _____

James G. Richards, Ph.D.
Vice Provost for Graduate and Professional Education

ACKNOWLEDGMENTS

Thank you to my family who supports me the most from every perspective through out my graduate academic education. Thank you to my advisor Hui Fang who offers me the opportunity to develop my idea further and supports me in many other ways. I am also grateful to all InfoLab members for their kind help. And to those not previously mentioned, who have influenced me or helped along the way.

TABLE OF CONTENTS

ABSTRACT

In this paper, we have addressed the problems of searching content in mathematical language, particularly measuring the similarity degree (in terms of structural and semantical) between mathematical expressions, summarized some general properties from mathematical semantics, that a search engine should be aware of. To better deal with these problems in an efficient way, we propose some ideas including: (1) A list of grammar rules to parse mathematical content (particularly in LaTeX markup) into a tree representation in order to preserve as much as information from mathematical expressions; (2) An index approach to break down the tree representation into what we call branch words to enable fast search in a similar fashion with inverted index, with parallelism potential; (3) A search method to capture some level of query-document subgraph isomorphism, combined with two pruning methods to both speed search and improve effectiveness. We also build our own proof-of-concept prototype search engine to demonstrate these ideas, and thus are able to present some evaluation results through this paper.

Chapter 1

BACKGROUND

Apart from general text content, structured information is also widely contained by digital document. Among these, a lot of mathematical content (including documents on Internet), are represented using markups like LaTeX or MathML [1], which is in a rich structural way. Information Retrieval on those structured data in mathematics language is not that well-studied or exhaustively covered by mainstream IR research, compared to that with general text. Thus it can be challenging yet very helpful given the contribution and importance of mathematics to our science.

However, the structured sense of mathematical language, as well as many its semantic properties (see section 1.2), makes general text retrieval models deficient to provide good search results. Through this paper, we have made our efforts to tackle some of these problems. Some of the ideas used in this paper deals with "tree structured" data in a general way, have the potential to be applied by other fields of structured data retrieval besides that from mathematical language.

1.1 Math IR Topics

Mathematical information involves a wide spectrum of topics, we are, of cause, not focusing on every aspects in mathematical information retrieval. It is good to clarify our concentration in this paper here, by first listing a set of concentrations that a mathematical information retrieval topics may be classified into, and define our target field of study.

[1] http://www.w3.org/Math/

Listed here, are considered four possible concentrations of topic for mathematical information retrieval:

1. Boolean or Similarity Search
2. Math Detection and Recognition
3. Evaluation, Derivation and Calculation
4. Other topics

The first one is doing mathematical information retrieval by searching, and finding the most relevant context of document that matches the query, very similar to the most common ways that other general text search engines will do, by boolean or similarity search [1]. The only difference is, the query may contain mathematical expressions. Instance of such search engine can be useful in many ways, for example, student may utilize it to know which identity can be applied to a formulae in order to give a proof of that formulae. This is the area where we focus in this paper. Specifically, we are proposing a series of methods for similarity search of math content. And our method is using query only in LaTeX markup (some math-aware search engines [2] support queries in mathematical formulae and normal text together), and return documents ordered by score which indicates the similarity degree.

Digital mathematical content document can also be in an image format (e.g. generated by a handwritten query), thus to retrieve these information involves detection or recognition. Inspired by the advances from deep learning, we may foresee a large potential to be explored on topics related to this.

Because the nature of mathematical language, a query (e.g. an algebra expression) can be evaluated and potentially derived into an alternate form, or calculated. The result value of evaluation or derived form may also be considered being relevant

[2] WolframAlpha: https://www.wolframalpha.com/ and Zentralblatt math from MathWebSearch: http://search.mathweb.org/zbl/

to that query. These potentially require a system to handle symbolic or value calculation, or even a good knowledge of derivation rules implied by different mathematical expression (e.g. computational engine *Symbolab* [3]).

Besides the first three concentrations, there are many other topics. Knowledge mining, for example, will need deeper level of understanding on math content. A typical goal of this topic is to give a solution or answer based on information retrieved from math content. e.g. "Find an article related to the *Four Color Theorem*" [2].

These concentrations somehow overlap in some cases, for example, some derivation can be used to better assess the similarity between math formulae, e.g. $\frac{a+b}{c}$ and $\frac{a}{c}+\frac{b}{c}$ should be considered as relevant. Therefore even boolean or similarity search possibly involves certain level of understanding of mathematics. In terms of similarity, however, we only address the measurement for structural and symbol differences in this paper, without considering further topics lured from measuring math content similarity, such as evaluation or derivation.

1.2 Issues in Measuring Similarity

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 these, we select and focus on those regarding to structural similarity and symbolic differences between expressions. At the same time trying to respect the semantical information inferred from structure or symbols in mathematical expressions. But even without caring about the possible derivations and high level knowledge inference, there are still many new problems.

Firstly, differences of symbols, structure and possible semantic rules in mathematics should be captured, and not one by one, but in an cooperative manner to measure similarity. To illustrate this point, we know that only respecting symbolic information is of course not sufficient in mathematical language. e.g. $ax+(b+c)$ in most

[3] Symbolab Web Search: http://www.symbolab.com

cases is not equivalent to $(a+b)x+c$ (although they have the same set of symbols). And the order of tokens in math expression can be commutative in some cases but not always. For example, commutative property in math makes $a+b=b+c$ for addition operation, but on the other hand $\frac{a}{b}$ is most likely not equivalent to $\frac{b}{a}$. These make many general text search methods (e.g. *bag of words* model, *tf-idf* weighting) inadequate. Moreover, symbols can be used interchangeably to represent the same meaning, e.g. $a^2+b^2=c^2$ and $x^2+y^2=z^2$. However, interchangeability comes with some constrains to maintain the same semantical meaning, 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 $a(1+b)$.

Secondly, how we evaluate structural similarity between expressions is a question. A complete query may structurally be a part of a document, or only some parts of a query match somewhere in a document expression. In cases when a set of matches occur within some measure of "distance", we may consider them to contribute similarity as a whole, but when matches occur "far away" for a query expression, then under the semantic implication of mathematics, they probably will not contribute the similarity degree in any way. We need a metres to score these similarity under certain criteria and set up standard and rules for relevance assessments.

Lastly, trying to capture semantic information from expressions will help measure similarity but introduce ambiguity. Apart from the cases covered in [3], semantic incorrect written markups, which is somehow common in many online documents, e.g. writing "sin" in LaTeX markup instead of macro "\sin", will make it difficult to tell whether it is a product of three symbols or a *sine* function, thus need to disambiguate. And depending on what level of semantical meaning we want to capture, ambiguity cases can be different. Consider $f(2x+1)$, if we want to know if f is a function rather than a variable, the only possibility is looking for implicit contexts, but we can nevertheless always think of it as a product without losing the possibility to search similar expression like $f(1+2y)$, the same way goes reciprocal a^{-1} and inverse function f^{-1}. Most often, even if no semantic ambiguity occurs, efforts are needed to capture

some semantical meanings. e.g. In $\int f(x)\frac{\mathrm{d}x}{\sin x}$ and $\sin 2\pi$, it is not easy to figure out, without a little knowledge on integral or trigonometric function, that integral is applied to $\frac{f(x)}{\sin x}$ and the scope applied by sine function is 2π, if we want to capture the subordinative relationship information.

1.3 Related Work

Boolean or similarity 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 [4]. And a variety of approaches have already emerged in a early timeline [5]. But there are a limited major ideas, from different angle, to deal with mathematical structured data. [6, 7, 8] use the same way to classify them into text-based and tree-based (structure-based). Here we follow the same classification and give a recap and an overview on their core ideas.

1.3.1 Text-based methods

Many researchers are utilizing existing models to deal with mathematical search, and use texted-based approaches to capture structural information on top of matured text search engine and tools (such as *Apache Lucene*).

DLMF project from NIST [9] uses "flattening process" to convert math to textualized terms, and normalize them into *sorted parse tree normal form* which creates a unique form for all possible orders of nodes (e.g. in a associative or commutative operator). Then further introduces serialization and scoping to stack terms [10], trying to capture structure information by using text-IR based systems that supports phrase search. Similar idea is also used by [11], expressions are also augmented for various possible representations, but variables are also replaced and normalized. The problem with augmentation in mathematical expression is obvious, complex expressions with many commutative operators will cost a lot of storage space, the benefits of capturing expression variances will be overshadowed.

The Mathdex search engine [12], from another perspective, uses query likelihood approach to estimate how likely the document will generate the query expression by a n-gram from root expression to sub-expression and tokens.

========== The conversion process loses considerable structural, and captures little semantics. There are approaches structure-based approaches [6] generate intermediate structure information

The MIaS system [13] try to reorder commutative operations and normalize variable and constance into unified symbols.

==========

Our system Cowpie [4] [**?**]

MathML vs LaTeX

distributed indexing to quickly search massive

Further more, a query may be specified with wildcards and thus will match any document with an expression substitution to that wildcard.

[4] demo page: infolab.ece.udel.edu:8912/cowpie/

REFERENCES

[1] Christopher D. Manning, Prabhakar Paghavan, and Hinrich Schutze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[2] 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.

[3] Richard J, Fateman, and Eylon Caspi. Parsing tex into mathematics. *SIGSAM Bulletin (ACM Special Interest Group on Symbolic and Algebraic Manipulation)*, 1999.

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

[5] Jozef Misutka. Mathematical search engine. Master's thesis, Charles University in Prague, May 2013.

[6] Xuan Hu, Liangcai Gao, Xiaoyan Lin, Zhi Tang, Xiaofan Lin, and Josef B. Baker. Wikimirs: A mathematical information retrieval system for wikipedia. *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries. Pages 11-20*, 2013.

[7] David Stalnaker and Richard Zanibbi. Math expression retrieval using an inverted index over symbol pairs. *Proc. SPIE 9402, Document Recognition and Retrieval XXII, 940207*, 2015.

[8] Qun Zhang and Abdou Youssef. An approach to math-similarity search. *Intelligent Computer Mathematics. International Conference, CICM*, 2014.

[9] Miller B. and Youssef A. Technical aspects of the digital library of mathematical functions. *Annals of Mathematics and Artificial Intelligence 38(1-3), 121136*, 2003.

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

[11] Jozef Miutka and Leo Galambo. Extending full text search engine for mathematical content. *Towards Digital Mathematics Library.*, 2008.

[12] Robert Miner and Rajesh Munavalli. *An Approach to Mathematical Search Through Query Formulation and Data Normalization.* Springer Berlin Heidelberg, 2007.

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

Appendix A

TITLE OF APPENDIX A

This is the information for the first appendix, Appendix A. Copy the base file, appA.tex, for each additional appendix needed such as appB.tex, appC.tex, etc. Modify the main base file to include each additional appendix file.

If there is only one appendix, then modify the main file to only use app.tex instead of appA.tex.

Appendix B

TITLE OF APPENDIX B