Enhancing Mathematics Information Retrieval

Martin Líška
Faculty of Informatics, Masaryk University, Botanická 68a, 602 00 Brno, Czech Republic
255768@mail.muni.cz

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information storage and Retrieval—*Information Search and Retrieval*; I.7 [**Computing Methodologies**]: Document and text Processing—*Index Generation*

Keywords

query expansion, digital mathematical libraries, indexing mathematics, MIaS, MathML, formula, subformula unification, evaluation

Motivation

Mathematics Information Retrieval (MIR) is a domain specific branch of Information Retrieval. MIR is a broad term for all the activities related to obtaining information from a collection of resources and answering an information need that involves mathematics in the form of math expressions and formulae. MIR is very important for Digital Mathematics Libraries (DMLs) which gather mathematically oriented documents in which users need to search and navigate effectively. A concrete implementation usually means a search engine that is able to answer a query composed of mathematical expressions as well as standard textual keywords searching through a collection with a substantial amount of mathematics.

There are several research groups that aim at creating well performing and usable math search engine. Our group located at the Faculty of Informatics, Masaryk University, develops Math Indexer and Searcher (MIaS) [2] – a math-aware search engine, currently deployed in EuDML (European Digital Mathematics Library). To the best of my knowledge, it is the only deployment of a MIR system in such a scale. Other deployments, e.g. DML-CZ, are planned.

The Goal

The goal is to design the ultimate math information retrieval system. The ultimateness can be understood as a combination of specific features that all together will support high precision as well as recall in searching documents with mathematical content.

Research Topics and Methods

The main research aspects of MIR are the following: *preprocessing* – unifying the input math notation; *indexing* – preparation of the internal structures and representation for similarity search; *searching* – interpretation of user query and expansion; *evaluation* – continuous evaluation of retrieval quality; *performance* – the internal representations must support a good overall performance.

Based on the NTCIR-11 Math-2 Task evaluation, MIaS approach is currently one of the best ranked in terms of the system's effectiveness [1]. In my work, I want to build upon the current performance of MIaS and research further methods and topics that could possibly make the system perform even better. The topics are the following:

Semantically Enhanced Search – integration of Computer algebra systems in the indexing as well as the searching phase as a computational power to reduce math expressions to a normalized form. Disambiguation of math structures from document and queries into known concepts in order to introduce semantics into the matching.

Subformula Unification – the ability to unify whole expression subtrees in order to allow the users to create queries containing wildcards in math expressions. In more complex query expressions wildcards can be critical in suppressing uninteresting parts of math expressions. The complexity of this aspect is heavily dependent on the representation of the data.

Query Expansion/Relaxation – mathematical queries can be relaxed by dividing it to smaller parts. This is can be a fallback method in case of unsuccessful complex query expressions. On the keyword level, mixed mathematical-textual queries can be relaxed by removing keywords from the individual parts of the query.

Combined Text-Mathematical Search – math expressions and text keywords need to be effectively combined for the system query in terms of logical operators as well as their mutual weighting.

Evaluation Framework – preparation of an automated evaluation framework for rigorously evaluating changes proposed to the retrieval process as listed above. This is possible with the presence of a test collection for MIR, such as NTCIR-11 Math-2 collection.

References

[1] M. Líška, M. Růžička, and P. Sojka. Math Indexer and Searcher under the Hood: History and Development of a Winning Strategy. In N. Kando, K. Kishida (Eds.), *Proceedings of the 11th NTCIR Conference*, pp. 127–134. NII, Tokyo, 2014.

[2] P. Sojka and M. Líška. Indexing and Searching Mathematics in Digital Libraries – Architecture, Design and Scalability Issues. In J. H. Davenport et al. (Eds.) *Proceedings of CICM 2011*, *LNAI 6824*, pp. 228–243, Berlin, Germany, 2011. Springer. `http://dx.doi.org/10.1007/978-3-642-22673-1_16`.

SIGIR'15, August 09-13, 2015, Santiago, Chile.
ACM 978-1-4503-3621-5/15/08.
DOI: http://dx.doi.org/10.1145/2766462.2767843.