SIMILARITY SEARCH FOR MATHEMATICS

MICHAL RŮŽIČKA, PETR SOJKA AND MARTIN LÍŠKA

THE WINNING STRATEGY OF THE NTCIR-11 MATH-2 TASK

INTRODUCTION

Math information retrieval (MIR) starts to be recognized as an important very domain-specific sort of information retrieval research field. Masaryk University (MU) has entered the area of MIR during the development of the Czech Digital Mathematics Library (DML-CZ) in mid-nineties. It became obvious that Digital Mathematical Libraries (DMLs) are specific in many aspects.

Masaryk University has partnered in the development of the European Digital Mathematics Library (EuDML) where it has been decided to support math formulae search as one of the math specific features. We have also paid attention to the user interface aspect: formulae in the query are rendered at the same time as the user writes it.

To the best of our knowledge, EuDML with Math Indexer and Searcher (MIaS) is the first digital library collecting non-born-digital PDFs that supports math search in full texts.

Our MIRMU team participated in the NTCIR-11 Math-2 math information retrieval task and compared the abilities of the MIaS system with other systems developed by researchers from around the world.

INDEXING AND SEARCHING

Our approach to searching mathematical content in documents is based on similarities of math structures through conventional full text searching. As mathematical notation, e.g. expressions and formulae, is highly structured we preprocess mathematical content in order to be processable by full text searching methods. The preprocessing procedures include canonicalization which is very important in order to allow matching of two equal formulae with slight notational differences. Therefore, the level of canonicalization needs to be as high as possible.

Then, to allow searching of subformulae, expressions are tokenized and subtrees of formulae extracted. Subformulae are stored in the locations of their original forms so they can be easily located at the query time. To be able to search for similar expressions we propose several generalization preprocessing techniques. These include unification of variables, unification of number constants, and font typeface preservation. These aim to increase the recall of mathematical search. To increase the precision we rank each indexed expression according to its distance from the original non-tokenized formula. The less unified subformulae extracted from a higher level of the original formula tree the higher weight factor it gets.

The very straightforward query interface of MIaS consists of only one input field. Users can type in textual queries together with math queries encoded using LaTeX notation as well as MathML notation. Queries are on-the-fly visualized as they are being 'typeset' in user's web browser to allow users to verify the correctness of the mathematical part of the query. Along the basic information about retrieved documents the result list shows a snippet with highlighted text and math tokens that are the most significant in the ranking of documents. This allows for quick primary evaluation of the document relevance to users' query.

Alongside interactive web querying interface MIaS offers searching using web services. This is a indispensable feature for automated querying that was used to retrieve evaluation results for the NTCIR Math Task.

AUTOMATIC QUERYING SCRIPTS

For NTCIR-11 Math-2 task evaluation we created a batch querying script that read topic specifications from the particular XML file and constructed four different XML queries for the MIaS web service interface for each of the topics:

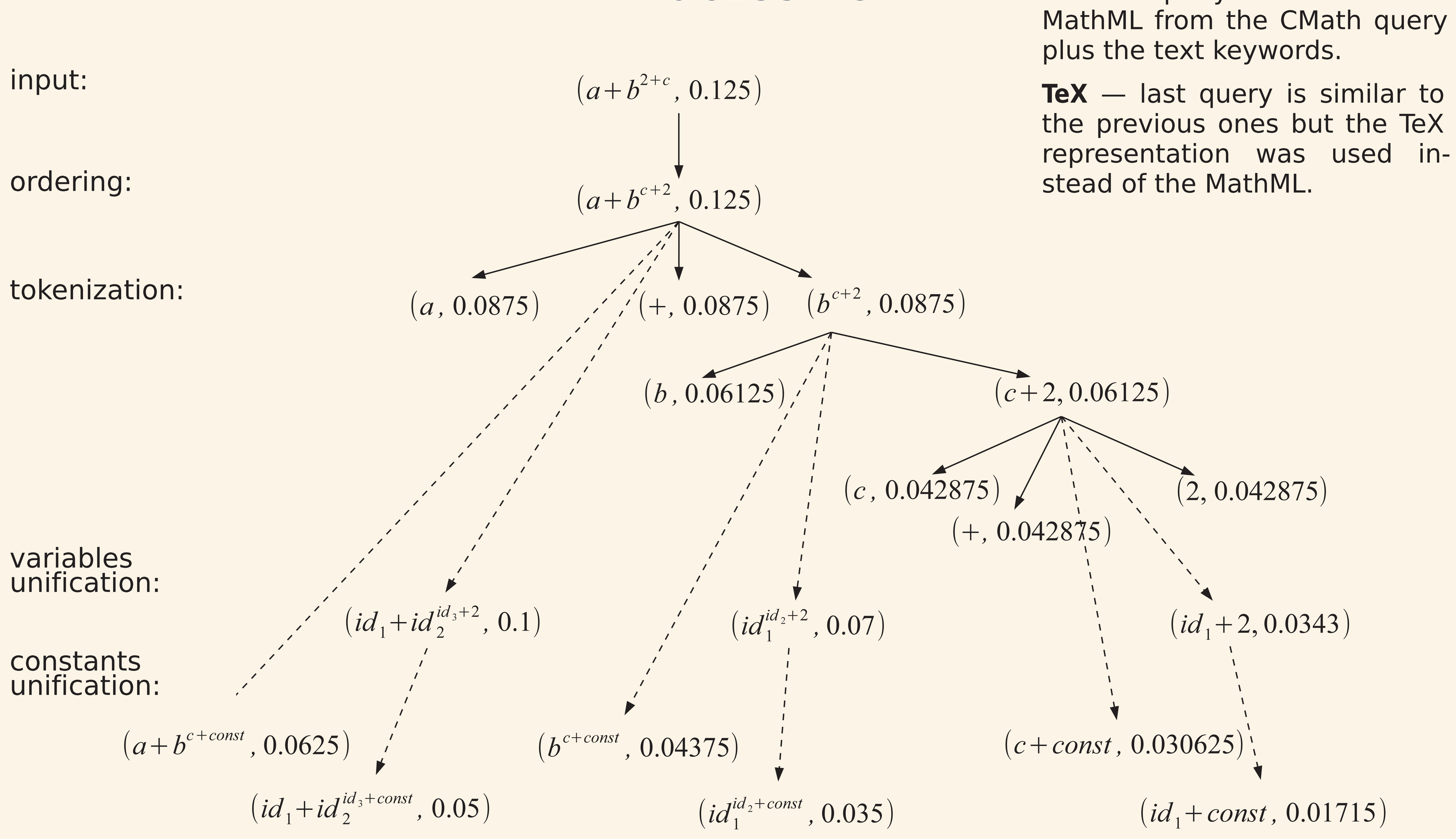
PMath — the query contained Presentation MathML representation of the query formulae together with text keywords.

CMath — this query was constructed in the very same way as the PMath query but using Content MathML representation of the formula instead of the Presentation MathML part.

PCMath — this query combines both Presentation and Content MathML, i.e. the query is constructed as concatenation of the Presentation MathML from the PMath query and Content MathML from the CMath query plus the text keywords.

TeX — last query is similar to the previous ones but the TeX representation was used instead of the MathML.

MATH PROCESSING



RESULTS

Table 1: Results of submitted runs with Relevance Level ≥ 3 (Relevant)

	PMath	CMath	PCMath	TeX	Rank
MAP avg	0.3073	**0.3630**	0.3594	0.3357	1
P-10 avg	0.3040	**0.3520**	0.3480	0.3380	1
P-5 avg	0.5120	**0.5680**	0.5560	0.5400	1

Table 2: Results of submitted runs with Relevance Level ≥ 1 (Partially Relevant)

	PMath	CMath	PCMath	TeX	Rank
MAP avg	0.2557	**0.2807**	0.2799	0.2747	2
P-10 avg	0.5020	0.5440	**0.5520**	0.5400	1
P-5 avg	0.8440	**0.8720**	0.8640	0.8480	2

Table 3: Index statistics

Indexing times [h]		Index
Wall Clock	CPU	size [GiB]
32.3	56.9	68

Table 4: Formulae count statistics

	Formulae	
Documents	Original	Indexed
8 301 545	59 647 566	3 021 865 236

EVALUATION

The result of all runs submitted by MIRMU team can be found in the tables on the left. The highest scores of our submitted runs highlighted in bold. The total rank of our best runs can be found in the last column — our system got the best results of all teams in 4 out of 6 evaluated categories but more importantly, we had the best score of all teams in the evaluation of results with Relevance Level ≥ 3.

PMath run based solely on Presentation MathML reached the lowest precision from our runs. Nevertheless, the average ratio between PMath run and CMath run raised from 0.64 in NTCIR-10 to 0.90 in the current evaluation. This is thanks to the constant development of our MathML Canonicalizer which is an important preprocessing step in the indexing as well as searching phase of search.

query 1 (the original query):	f_1	f_2	k_1	k_2	k_3
query 2:	f_1	f_2	k_1	k_2	
query 3:	f_1	f_2	k_1		
query 4:	f_1	f_2			
query 5:	f_1		k_1	k_2	k_3
query 6:			k_1	k_2	k_3

Example 1: Complete sequence of subqueries derived from the original user's query.

QUERY EXPANSION

Combination of multiple formulae and multiple text keywords in one query used in NTCIR-11 Math Task seems to be more consistent with the real situation of a human using maths-aware search engine: formulae are simply a different expression of keywords used to filter relevant documents from the whole database. They are complement instrument of the query specification to the keywords, not the opposite of them. The queries work best with formulae and keywords together.

The MIaS system supports this kind of queries natively. All the parameters are posted to the system in one text field — formulae are written in MathML or TeX notation with a dollar sign ($) added on both sides of the TeX formulae. Keywords, sometimes consisting of more than one word, were surrounded with single quotation mark (") to handle multi-word keywords as a single entity. Formulae and text keywords were separated by a single space.

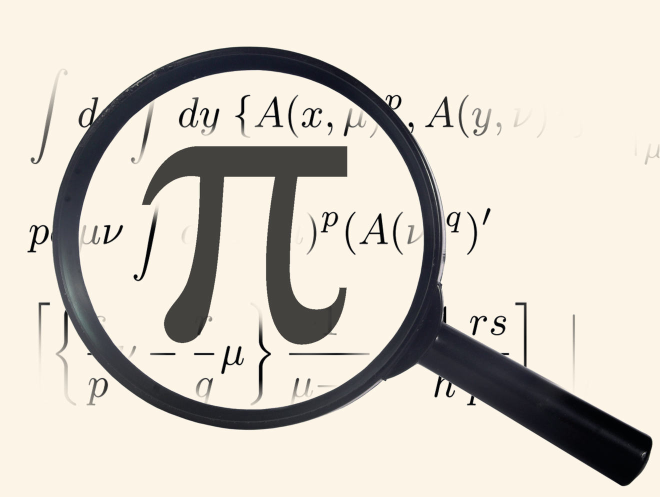
Internally, we are expanding the original query further to increase recall on very specific queries with no or just minimal number of results found.

RESULTS MERGING

Every subquery results in an ordered list of results with score assigned to each of the results. However, these scores are only comparable within the context of the results list. Thus, it is not possible to generate final results list as a simple combination of results from all the subqueries ordered by the score.

Another reason to use a more complicated results merging procedure is a necessity of preference of results on the original user's query to the results found for subqueries. On the other hand, it is well possible that the first result of a subquery could be more relevant for the user than the 153th result on the original query.

To produce the final results list from the subqueries according to this hypothesis we used a method we refer to as 'strip-merging' of the results. The main idea is interleaving of 'strips' of results from all the ordered results lists from the subqueries. The less modified subquery to the original query the 'wider' strip of results on more relevant position is used in the final result list.



CONCLUSION

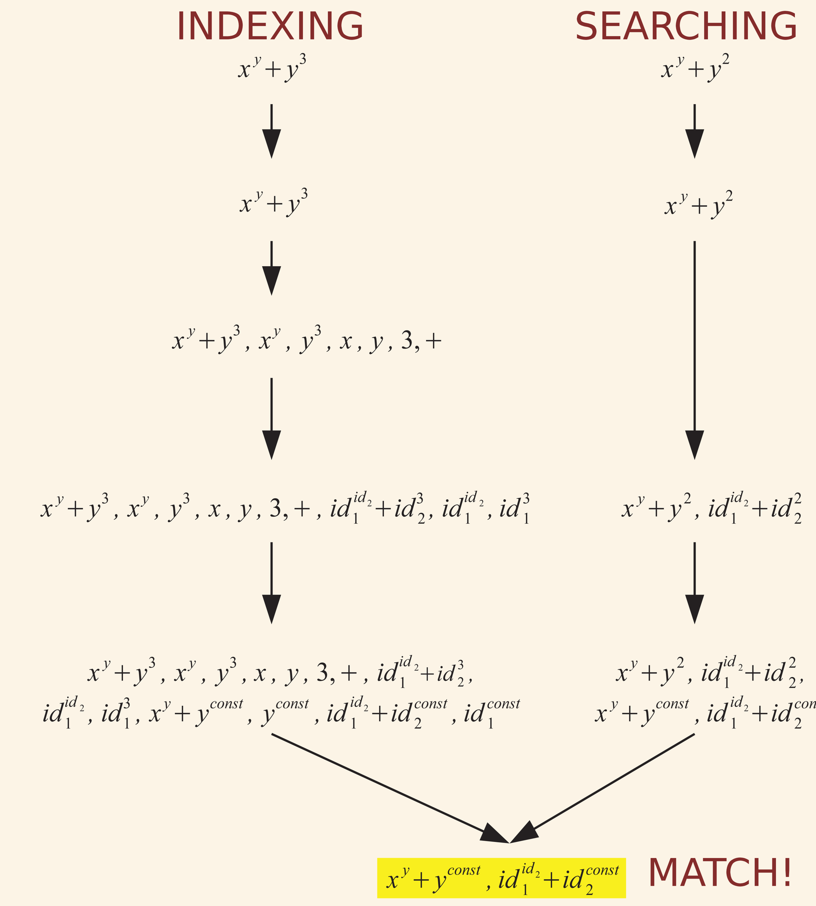
Our participation in the NTCIR-11 Math 2 Task was very useful and motivating for the development of our system. It provides unique opportunity to directly compare different systems with different approaches on the same data set. Our MIaS system achieved best results in the targeted subtasks.

We would like to inspect on every judged results that we posted to understand its non/relevancy to the respective query. The MIaS system matches every formula in the same manner and yet, there are results returned by our system some of which were found relevant also by the judges and some of which were found non-relevant. We hope to find patterns in what makes some results non-relevant which could lead us to improving our system.

Further investigation of the best strategies of subqueries derivation from the original users' query is needed as well as proper evaluation of different strategies of merging results of subqueries to the final result list.

As for future work we would like to focus on improving formula unification techniques. Specifically, complete formula subtree unification can improve searching based on formula structure.

WORKFLOW EXAMPLE



Contact Information

mruzicka@mail.muni.cz, sojka@fi.muni.cz, martin.liski@mail.muni.cz, Faculty of Informatics, Botanická 68a, 602 00 Brno, Czech Republic

https://mir.fi.muni.cz/

Citation

RŮŽIČKA, Michal, Petr SOJKA and Martin LÍŠKA. Math Indexer and Searcher under the Hood: History and Development of a Winning Strategy. In Hidei Joho, Kazuaki Kishida. Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies. Tokyo: National Institute of Informatics, 2-1-2 Hitotsubashi, 2014. 8 pp. ISBN 978-4-86049-065-2.

Acknowledgements

We acknowledge the support (Short and Exchange Visit Grants 6965 and 6967) received from ESF, European Science Foundation, for the activity entitled ELIAS—Evaluating Information Access Systems.

FACULTAS ARTIS INFORMATICÆ • UNIVERSITAS MASARYKIANA •

Publications

SOJKA, Petr and Martin LÍŠKA. The Art of Mathematics Retrieval. In Matthew R. B. Hardy, Frank Wm. Tompa. Proceedings of the 2011 ACM Symposium on Document Engineering. Mountain View, CA, USA: ACM, 2011. p. 57–60, 4 pp. ISBN 978-1-4503-0863-2. doi:10.1145/2034691.2034703.

SOJKA, Petr and Martin LÍŠKA. Indexing and Searching Mathematics in Digital Libraries – Architecture, Design and Scalability Issues. In James H. Davenport, William M. Farmer, Josef Urban, Florian Rabe. Intelligent Computer Mathematics Lecture Notes in Computer Science, 2011, Volume 6824/2011. Berlin / Heidelberg: Springer, 2011. p. 228–243, 15 pp. ISBN 978-3-642-22672-4. doi:10.1007/978-3-642-22673-1_16.

SOJKA, Petr and Martin LÍŠKA and Michal RŮŽIČKA. Building Corpora of Technical Texts : Approaches and Tools. In Aleš Horák, Pavel Rychlý. Fifth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2011. Brno: Tribun EU, 2011. p. 71–82, 11 pp. ISBN 978-80-263-0077-9.

LÍŠKA, Martin and Petr SOJKA and Michal RŮŽIČKA and Peter MRAVEC. Web Interface and Collection for Mathematical Retrieval : WebMIaS and MREC. In Petr Sojka, Thierry Bouche. DML 2011: Towards a Digital Mathematics Library. Brno: Masaryk University, 2011. p. 77–84, 8 pp. ISBN 978-80-210-5542-1.