

Similarity Search for Mathematics: Masaryk University team at the NTCIR-10 Math Task

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

Petr Sojka
Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno
Czech Republic
sojka@fi.muni.cz

Michal Růžička
Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno
Czech Republic
mruzicka@mail.muni.cz

ABSTRACT

This paper describes and summarizes experiences of Masaryk University team MIRMU with the mathematical search performed for the NTCIR pilot Math Task. Our approach is the similarity search based on enhanced full text search utilizing attested state-of-the-art techniques and implementations. The variability of used Math Indexer and Searcher (MIaS) system in terms of the math query notation was tested by submitting multiple runs with four query notations provided. The analysis of the evaluation results shows that the system performs best using \TeX queries that are translated to combined Presentation-Content MathML.

Team Name

MIRMU (Math Information Retrieval at Masaryk University)

Subtasks

Math Retrieval Subtask

Keywords

math, search, similarity search, math information retrieval, MIR, MIaS, evaluation, math representation and indexing

1. 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. This fact motivated and triggered establishment of DML workshop series in 2008 [7].

Some papers in DMLs consist of more formulae than texts, and we started to think about representation and indexing of mathematical formulae *in addition* to texts. Because formulae appearing even in the main metadata (title, abstract, references) were not properly recognized, represented, and indexed, handling of math in Google Scholar or DMLs was notably suboptimal. There was no widely acceptable user interface and representation for math formulae in information retrieval (IR). \TeX math has been designed for typesetting and optimized for minimal stroke typing. Luckily, logical math markup of \LaTeX is widespread now, and AMS packages cover most of math needed in IR.

We have designed and developed first math formulae indexing and retrieval prototypes in the series of Bachelor thesis [2, 5]. Math formulae are structures appearing within accompanying texts that convey meaning and relations between objects mentioned in the text. They could be represented as *trees* and one could define *formulae similarity as tree structure similarity*.

The working prototypes were further experimented with, developed, researched [9, 10], discussed at DML panels and evaluated [6]. First MIR specific workshop was co-organized in Bremen in 2012, to heat the discussions about MIR as a gateway to the vast knowledge stored in DMLs.

MU has partnered in the development of the European Digital Mathematics Library, EuDML [11], where it has been decided to support math formulae search, as one of math specific features. [12] We have also paid attention to the user interface aspects—formulae is rendered as user types by rendering the formulae after every keystroke [13, Section 3.5]. To the best of our knowledge, EuDML with MIaS is the first digital library collecting non-born-digital PDFs that supports math search in fulltexts. Mathematician all over the world can use it and practise math search as a mean to narrow their search with math formulae facets.

The paper is structured as follows. In the Section 2 we give a brief overview of our approach used in our MIaS system. Section 3 describes run that MIRMU team submitted for NTCIR Math tasks, with scripts used for automation of querying. In the Section 4 our approach in all three participated subtasks is revealed. Results are discussed in the Section 5. We conclude with summary and further development thoughts in the Section 6.

2. OVERVIEW

Our approach to searching mathematical content in documents is based on 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 [3]. 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 preprocess-

ing 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. Assigned weights affect the ordering of retrieved results.

The factors that influence resulting weights of indexed subformulae are adjustable. Different document collections, i.e. from different STEM fields, benefit from different setups. The current setup reflects our generic view of distance of extracted subformulae to their original trees. Different setup might influence the order of retrieved results significantly. From the evaluation point of view, different setups bring different hits to the pooling resulting in higher or lower relevancy. There is no ideal set of factors, however, we want to reach to the optimal setup by repetitive evaluation as discussed in the Section 6.

We developed a search system according to these principles. MIaS (Math Indexer and Searcher) is a math-aware full-text based search engine. It is based on the state-of-the-art searching library Lucene. It supports combined text and math searching. Refinement of many text query results by adding a math query is believed to be a very powerful tool. Mathematical preprocessing is a plug-in that can be used with any Lucene or Solr based systems. This is the case of EuDML. MIaS processes documents with mathematics encoded in Presentation or Content MathML. At the end of the preprocessing, expression trees are linearized to compacted string form to reduce index space requirements.

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. Query is on-the-fly visualized as ‘typeset’ formula 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 document’s rank. This allows for quick primary evaluation of the documents relevance to user’s query.

Alongside interactive web querying interface MIaS offers searching using web services. This is an indispensable feature for automated querying that was used to retrieve evaluation results for the NTCIR Math Task. [1]

MIaS participated in the MIR 2012 Happening evaluation workshop with good performance. Further evaluation for both effectiveness and efficiency using MIR Happening as well as NTCIR Math Task collections was done in [6].

For a more detailed description of all parts of the system the reader is referred to [9, 8, 6].

3. AUTOMATIC QUERYING SCRIPTS

Availability of the task inputs in the XML format supplemented by MIaS web service interface allowed us to fully automatize the task data processing.

For the Formula Search and Full Text Search subtasks the batch querying script read subtask topic specifications from the particular XML file and constructed four different XML queries for the MIaS web service interface for each of the subtask topics:

PMath query The query contained Presentation MathML code specified in the `pquery` element of the subtask topic XML specification. In the case of the Full Text Search subtask, plain text from the `words` element of the particular topic was added to the end of the query. The input MathML code was slightly modified—each occurrence of the construction

```
<mws:qvar xmlns:mws="http://www.mathweb.org/mws/ns"
          name="X"/>
```

was substituted for a simple `<mi>X</mi>` statement.

CMath query This query was constructed in the very same way as the PMath query but using Content MathML, i.e. `cquery` element was used instead of the `pquery` element and `qvar` was substituted for the `ci` element.

PCMath query 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, in the case of the Full Text Search subtask followed by the plain text from the `words` element.

\TeX The last query is similar to the previous ones but the \TeX code from the `TeXquery` element was used instead of the MathML. There was no modification of the \TeX statement except for a single dollar sign (\$) added on both sides of the original statement to properly indicate \TeX encoded part of the query to the MIaS system.

Presentation MathML is widely used XML encoding of mathematical contents and is often used together with the XHTML markup language on the web. It is also used as a common format for data interchange among various computer systems. Encoding just the *appearance* of the formulae it is reasonably easy to generate Presentation MathML from other languages, e.g. \TeX . Being widely used we believe a math-aware search engine have to be able to cope well with this MathML encoding.

On the other hand, Content MathML encodes *meaning* of the formulae. With proper normalization of the input encoding, it should be easier to find similar formulae using this form of MathML. Through the CMath and PCMath queries we wanted to compare behaviour of our system in contrast to the query using Presentation MathML notation.

The \TeX query was used to investigate impact of the \LaTeX to MathML on-the-fly conversion that is performed by the MIaS system if the user ask queries in this ‘human-friendly’ language.

For the Open Information Retrieval subtask we had to prepare MIaS queries based on the natural text specification of the problem. Queries were constructed as set of keywords for each topic usually extended by mathematical formula in the \TeX notation. Technically, query was constructed similarly to the \TeX queries for the Full Text Search subtask. To allow automation of the subtask processing our query data was saved in a simple machine processable plain text format that was read by the batch script.

The constructed XML query was sent to the MIaS web service. Two different indexes were used—for the Formula Search subtask index for finding and retrieving single formulae was used, the Full Text Search and Open Information

Retrieval subtasks used standard MIaS index for document finding (see Section 4).

MIaS web service answered in a XML format. The answer was analyzed by our script and three different output formats were created.

text The text format was constructed according to the NTCIR Math Task requirements for the text results submission format. Topic ID, found document/formula ID and its rank (order in the result list) together with MIaS score assigned to the result and identifier of the run was presented in a simple plain text format.

XML The NTCIR Math Task defines also XML format for submission of the results. This file encodes in XML similar information as the text format and was used for final submission of the MIRMU results.

HTML For investigation of the results during the tuning of the system and queries we generated HTML summary of the results. In this format, plain text representation of the results in the NTCIR Math Task format was shown. However, in addition to that hyperlinks for direct navigation to the WebMIaS interface were available for interactive investigation of the queries and results together with MIaS web service response XML. See Figure 1.

The querying process of all nine runs we performed for participation in the NTCIR Math Task took less than one hour (roughly 6 minutes each) of clock time on a standard PC workstation.

4. MATH RETRIEVAL SUBTASK

MIRMU team participated in the Math Retrieval Subtask with contributions to all three types of search: Formula Search, Full Text Search and Open Information Retrieval.

Full Text Search simulated the standard use of a search system—queries comprised math expressions as well as text. For each query, the system returned a list of documents as they were provided in the test collection. No special modifications were therefore needed.

For the Formula Search, however, several adjustments were necessary. Formula Search aimed at retrieving independent formulae located in the provided documents. If, for example, a document contained 100 formulae, each of them could be retrieved as a hit on its own. This is a difference to the normal workflow. However, flexible design of MIaS allowed us to index every formula as an independent index document containing only that formula by adding a special document handler.

For the needs of Math Retrieval Subtask, we created two indexes from the provided document collection, that contained 36,697,971 math expressions and had 7.3 GB in size. After preprocessing, both indices stored more than 1.5 billion subexpressions. The first index, NTCIR-fragments, was created from single formulae to complete Formula Search type. Every index document represented only one formula from the input files, therefore, the resulting index contained more than 73.5 million documents. It took 8.5 hours to complete the index sized around 39.5 GB. The second index called NTCIR-files was created the regular way consisting both of text and formulae where one index document represented exactly one physical document from the collection. It took 5 hours to complete the index sized around

30 GB. This comparison shows an interesting overhead of the Formula Search index. It contains less data but is split into more logical units which resulted in the longer indexing time and a larger index.

Table 1: Index statistics (run on 448GiB RAM, eight 8-core 64bit processors Intel Xeon TM X7560 2.26GHz machine)

Index	Indexing times [min]		Index size [GB]
	Wall	CPU	
NTCIR-files	291.8	1649.0	30
NTCIR-fragments	513.3	2029.4	39.5

Alongside text, MIaS accepts \LaTeX and both Content and Presentation MathML as a query notation for mathematics. \LaTeX queries are converted to combined Presentation-Content MathML by \LaTeX XML converter. We decided to utilize the possibility of submission of four runs to analyse the difference in the performance of the system with regard to the query language. This was supported by the test query collection that provided all of the mentioned formats for each query. Table 2 shows differences in query language for each run.

Table 2: Runs submitted to Formula Search and Full Text Search

Run #	Query language
1	Presentation MathML
2	Content MathML
3	Presentation and Content MathML
4	\TeX

After the results publication we discovered that Run 2 and Run 3 produced exactly the same hit lists with the same results. Therefore, we omit erroneous Run 3 from our results analysis in Section 5. We estimate its effectiveness to be at the same level with the \TeX Run 4.

Open Information Retrieval Subtask

Open Information Retrieval subtask is probably closest to the real use and setup of MIR both in digital library systems or on the web. People are used to query Google just by giving small set of keywords, and find what they are looking for. Mathematicians are used to write their formulae in \LaTeX are capable of specifying domain and semantically related formulae in a query.

We took the advantage of textual index of our system—MIaS has both text and math indexed in a Lucene based system. Queries were written manually mainly as bag of text and formulae. Textual area allows querying of alphabetical or alphanumerical words, or collocations, and search was narrowed usually by additional bag of formulae.

The second author has prepared a set queries in a text file, from which the queries containing all text and MathML (both Presentation and Content, by LaTeXXML) were generated by a script. The strategy was to find and write down the biggest and ‘semantically’ close set of formulae and words. As MIaS scoring is cumulative, the more, the better, given the laziness of user. Examples from a submitted query file is below (query NTCIR code is follow by line of formulae and line of textual keywords and collocations:

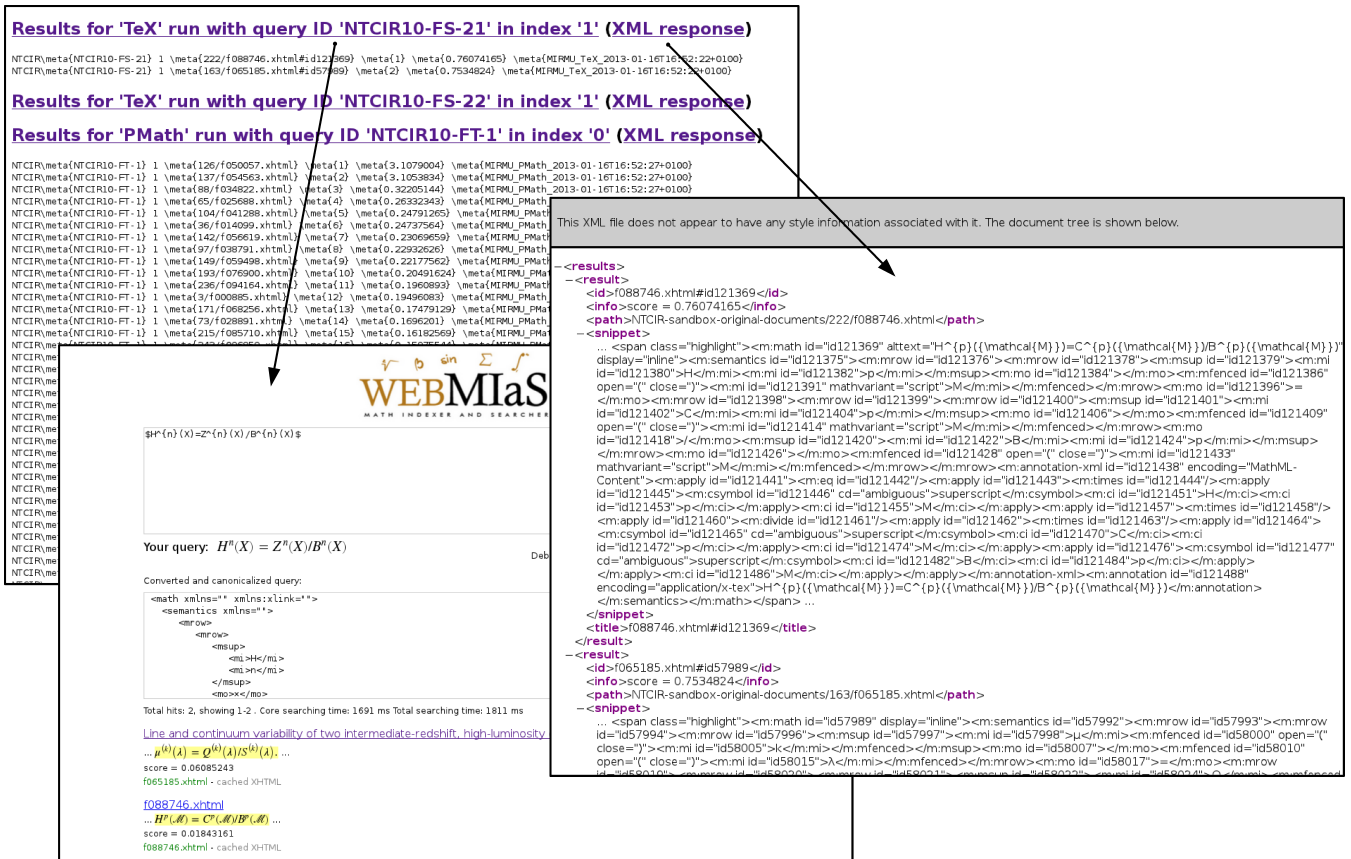


Figure 1: HTML summary for the investigation of the results.

- NTCIR10-OMIR-1
 - $\$p \in (1, \infty)$ $\$L_1(L_p)$ $\$L_q$
 - conjugate space isomorphic subspace
 - NTCIR10-OMIR-2
 - $\$\Gamma \text{ Hilbert space differential operator discrete amenable group}$
 - Neumann algebra
 - NTCIR10-OMIR-3
 - $\$\Delta$ differential equation operator
 - NTCIR10-OMIR-4
 - $\$y'' + p(x)y' + q(x)y = f(x)$ $\$F : R^n \rightarrow R^n$
 - second order vector continuous "bounded solution"
 - "differential equation"
 - NTCIR10-OMIR-5
 - $\$\implies$
 - reduction ad absurdum
 - NTCIR10-OMIR-6
 - four color theorem
 - NTCIR10-OMIR-7
 - associative operators quantum mechanics
 - NTCIR10-OMIR-8
 - commutative monoid
 - NTCIR10-OMIR-9
 - perfect graph theorem Lovász
 - NTCIR10-OMIR-10
 - $\$\implies$
 - proof by induction
 - NTCIR10-OMIR-11
 - Nonlinearity Predicting Chaotic System
 - NTCIR10-OMIR-12
 - sequentially compact "limit point compact" compact
 - totally bounded complete
 - NTCIR10-OMIR-13
 - 40A05
 - NTCIR10-OMIR-14
 - NTCIR10-OMIR-15
 - Riemann theta function
 - NTCIR10-OMIR-16
 - $\$ax^3 + bx + c$
 - Cardano cubic equation
 - NTCIR10-OMIR-17
 - $\$n^5 - 7n^4 + 17n^3 - 18n^2 + 7n - 1$
 - root polynomial
 - NTCIR10-OMIR-18
 - zero divisor invertible finite ring inverse is power
 - NTCIR10-OMIR-19
 - $\$f : R^3 \rightarrow R$
 - function derivative real-valued analysis
 - nonempty closed set intersection Banach space
- For example, for the OMIR-2 query we have used as many words and formulae as we thought could appear in a relevant document. Another example worth mentioning is query OMIR-13 where alphanumerical token of Mathematical Subject Classification code was used as the only element of the query: '40A05'.
- In fact, preparation of the queries was straightforward with just few debugging queries used during the development process, especially when one got no results.

5. RESULTS

Overall scores of MIaS were above average of the Math Task results [1]. Precision at rank five (P-5) of MIaS in Run 4 was the highest from the all competing submissions. Table 3 shows all four reported metrics for relevance level ‘relevant’. Table 4 shows the same metrics for relevance level ‘partially relevant’.

Interestingly, while other systems roughly doubled their scores in ‘partially relevant’ metrics compared to the ‘relevant’ ones (see the overview paper [1] for all results), results of MIaS increased only slightly. We find this behavior very interesting, as our system is a *math similarity* search system, therefore we would expect ‘partially relevant’ results to be much more successful.

We think this is caused by the design of MIaS that does not tokenize queries. The complexity of the queries was relatively high and MIaS does not extract subexpressions from queries and so it searches only for the whole query formulae. Therefore, when the query is complex, it may match it as a whole or not at all.

We do not know the other systems in detail, but another possible explanation for this behaviour can be different level of unification done by the MIaS system compared to other systems. We do match only formulae with the same structure, but there are other systems that do either full unification or also weight and match formulae with different or bigger structure. The unification in MIaS is done on the leaf level of derivation trees of formulae, namely unification of variables and unification of number constants. No unification on the inner node level is done so far, therefore MIaS is unable to substitute subtrees. For example, $\sum_{i=1}^n$ will not

match $\sum_{i=n-5}^n$ as $(n - 5)$ is a different subtree to 1.

Another use case is when searching small formulae. In MIaS, every subcomponent of a complex formula is indexed. Thus, even frequent simple formulae queries such as x^2 that occur in a large fraction of all the documents contributes slightly to the partial relevance of each of these documents. Therefore, almost every document is partially relevant and the discrimination between documents is small. Possible remedy is to use equivalent of inverse document frequency scheme (IDF) weighting for math structure (size)—inverse math structure document frequency (IMSDF): the more frequent formulae [structure] occur in the documents the less weight it will get during indexing, e.g. the weight will be multiplied by $\log \frac{\# \text{ of all documents}}{\# \text{ of structurally same formulae}}$.

Both tables show that the most successful run was Run 4 that used \TeX queries. Precision was the best in Run 2. Run 4 retrieved 50 % more results than Runs 2 but only few of these additional items were judged as relevant or partially relevant.

Table 3: Result metrics for submitted runs in Formula Search with Relevance Level ≥ 3 (Relevant)

Metric	Run 1	Run 2	Run 4
P-10 avg	0.105	0.191	0.219
P-5 avg	0.133	0.229	0.276
MAP avg	0.060	0.112	0.127
Precision	0.109 (64/589)	0.185 (92/496)	0.123 (96/778)

Table 4: Result metrics for submitted runs in Formula Search with Relevance Level ≥ 1 (Partially relevant)

Metric	Run 1	Run 2	Run 4
P-10 avg	0.143	0.214	0.267
P-5 avg	0.181	0.267	0.343
MAP avg	0.066	0.081	0.100
Precision	0.148 (87/589)	0.232 (115/496)	0.161 (125/778)

Unfortunately, there has been a cardinal misunderstanding about what the results of Full Text and OMIR searches should report. Our search system for mathematics MIaS and, to our knowledge, the majority of standard search engines are designed to report results at the document level. With MIR system, one can search only with formulae, but still expects whole documents to be returned as results. This becomes even more obvious, when math query is combined with text to make the query more precise which, we believe, is the right way of making use of math search capability. This is exactly the case of Full Text and OMIR search types. Queries consisted of both text and math which exactly simulates the right usage of MIR systems. We do not understand how only formula references can justify results of mixed math-textual query. Additionally, as our system is based on full-text search, we consider the implicit combination of math and text search with the powerful weighting function as one of our biggest advantages of our system. On the other hand, we understand the severity of evaluating the whole documents as results.

Nevertheless, our team posted document identifications as results in both Full Text and OMIR searches which made them invalid. Our system therefore could not be evaluated in these natural search types.

6. CONCLUSIONS

Our participation in the NTCIR pilot Math 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 satisfactory 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.

The Math Pilot Task detected the best query language for our system with the respect to the effectiveness of the system. It is a \TeX notation that is translated to MathML parallel Presentation-Content markup. This is gratifying news as we also want \TeX to be used as the main query language since it is much used by the scientists and therefore well known. To encourage \TeX input to the WebMIaS (web interface of our system), we added the on-the-fly rendering of the input formula for a better user experience and the verification of correctness of the formula.

We discovered, that ability to evaluate is very valuable in information retrieval. It is a driving force in the evolution process of IR systems, more so if it is impartial as for example at the NTCIR conference task. But, to justify the development on a day to day basis, we need our own collection with gold standards against which we could evaluate our development steps. All of the participants as well as the organizers of the Math Pilot Task recognized the difficulties connected to evaluating math search systems. Our future goal is to create our own, gold standard evaluation collection. We find it a prerequisite to the further development of retrieval techniques.

Future Work

We established the best query language for mathematical part of queries. In the following Math Task, and we hope there will be one, we will elaborate more on the weighting function and make use of the runs by submitting results generated with different factors that are used for the computation of the similarity between the original formula and its derived forms. This can shuffle the order of search results to make more relevant hits appear higher in the result list.

Some ideas and aims of the future development of MIaS and related tools are summarized in [8]. We believe that the way to increase the practical usability and success of a MIR system like MIaS is conditioned by better *relevance* and *speed* (users love instant feedback). We plan to experiment with several tier indexing for speed and prioritization of semantics (Context MathML). A prerequisite for increasing the relevance is better formulae *disambiguation* and *canonicalization*. The road to it is paved by the full natural language processing of the corpora of math texts (part of speech tagging, named entity recognition, document classification) and further method adapted for math from corpus linguistics done on texts as methods for computing *semantic relatedness*. This would allow us to take participation in the Math Understanding Task next NTCIR. In this context and in long term we expect to experiment with *Explicit Semantic Analysis* (ESA) [4] adapted to MIR using also math formulae, connotations and named entities to model semantic relatedness. We plan to measure its impact on MIR qualities on a developed reference document and query corpus for MIR evaluation.

Acknowledgement. We acknowledge the support (Short and Exchange Visit Grant) received from the European Science Foundation (ESF) for the activity entitled ‘Evaluating Information Access Systems’.

7. REFERENCES

- [1] A. Aizawa, M. Kohlhase, and I. Ounis. NTCIR-10 Math Pilot Task Overview. In *Proceedings of the 10th NTCIR Conference*, Tokyo, Japan, 2013. To appear.
- [2] V. Dostál. Indexing of Mathematical Texts in the Digital Mathematics Library (in Czech), Jan. 2009. Master Thesis, Masaryk University, Brno, Faculty of Informatics (advisor: Petr Sojka), https://is.muni.cz/th/72569/fi_m/?lang=en.
- [3] D. Formánek, M. Líška, M. Růžička, and P. Sojka. Normalization of digital mathematics library content. In J. Davenport, J. Jeuring, C. Lange, and P. Libbrecht, editors, *24th OpenMath Workshop*, *7th Workshop on Mathematical User Interfaces (MathUI), and Intelligent Computer Mathematics Work in Progress*, number 921 in CEUR Workshop Proceedings, pages 91–103, Aachen, 2012. <http://ceur-ws.org/Vol-921/wip-05.pdf>.
- [4] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI '07*, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc. <http://dl.acm.org/citation.cfm?id=1625275.1625535>.
- [5] M. Líška. Searching Mathematical Texts (in Slovak), May 2010. Bachelor Thesis, Masaryk University, Brno, Faculty of Informatics (advisor: Petr Sojka), https://is.muni.cz/th/255768/fi_b/?lang=en.
- [6] M. Líška. Evaluation of Mathematics Retrieval, Jan. 2013. Master Thesis, Masaryk University, Brno, Faculty of Informatics (advisor: Petr Sojka), https://is.muni.cz/th/255768/fi_m/?lang=en.
- [7] P. Sojka, editor. *Towards a Digital Mathematics Library*, Birmingham, UK, July 2008. Masaryk University. <http://www.fi.muni.cz/~sojka/dml-2008-program.xhtml>.
- [8] P. Sojka. Exploiting Semantic Annotations in Math Information Retrieval. In J. Kamps, J. Karlgren, P. Mika, and V. Murdock, editors, *Proceedings of ESAIR 2012 c/o CIKM 2012*, pages 15–16, Maui, Hawaii, USA, 2012. Association for Computing Machinery. <http://doi.acm.org/10.1145/2390148.2390157>.
- [9] P. Sojka and M. Líška. Indexing and Searching Mathematics in Digital Libraries – Architecture, Design and Scalability Issues. In J. H. Davenport, W. M. Farmer, J. Urban, and F. Rabe, editors, *Proceedings of CICM 2011*, volume 6824 of *LNAI*, pages 228–243, Berlin, Germany, July 2011. Springer-Verlag. http://dx.doi.org/10.1007/978-3-642-22673-1_16.
- [10] P. Sojka and M. Líška. The Art of Mathematics Retrieval. In *Proceedings of the ACM Conference on Document Engineering, DocEng 2011*, pages 57–60, Mountain View, CA, Sept. 2011. Association of Computing Machinery. <http://doi.acm.org/10.1145/2034691.2034703>.
- [11] W. Sylwestrzak, J. Borbinha, T. Bouche, A. Nowiński, and P. Sojka. EuDML—Towards the European Digital Mathematics Library. In P. Sojka, editor, *Proceedings of DML 2010*, pages 11–24, Paris, France, July 2010. Masaryk University. <http://dml.cz/dmlcz/702569>.
- [12] K. Wojciechowski, A. Nowiński, J. Grimley, and M. Líška. Public User Interface – Final Release, Jan. 2013. Deliverable D6.5 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library.
- [13] K. Wojciechowski, A. Nowiński, P. Sojka, and M. Líška. The EuDML Search and Browsing Service - Final, Jan. 2013. Deliverable D5.3 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, revision 1.1 <https://project.eudml.eu/sites/default/files/D5.3-v1.1.pdf>.