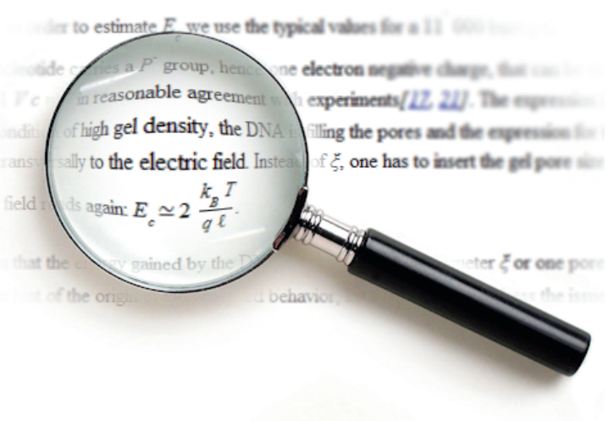
Martin Líška, Petr Sojka and Michal Růžička

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

Similarity Search for Mathematics

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



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.

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. There was no widely acceptable user interface and representation for math formulae in information retrieval (IR). We have designed and developed first math formulae indexing and retrieval prototypes in the series of Bachelor thesis. 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.

MU 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 math specific features. We have also paid attention to the user interface aspects: formulae is rendered as user types by rendering the formulae after every keystroke.

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

Canonicalization

Full paper texts have to be ‘homogenized’, converted to some uniform representation, in order for math-aware fulltext searches and paper similarity computations to work properly.

There is an average of 380 mathematical formulae per arXiv paper in our MREC database. It has been reported that even a single histogram of mathematical symbols is sufficient for domain classification of a paper in the mathematical domain. To reliably represent a paper for DML processing, including handling the mathematics, it is necessary to

1. select a canonical representation of the non-textual structural entities appearing in fulltexts (mathematical symbols, formulae, and equations); and
2. decide on equivalence classes for these entities (e.g., for which formulae should be considered equal for given DML tasks such as search, similarity computation, formulae editing, and conversion of math into Braille).

Using our public working demo of the WebMIaS system we discovered several discrepancies in the form of MathML generated by the real-time TeX to MathML converter we currently use – Tralics. We tried to normalize the users’ MathML input and the MathML produced by the LaTeXML converter contained in the arXMLiv collection. Then we went through the Presentation MathML specifications and gathered a list of possible reformatting rules we could perform.

Sub-/Superscripts Handling

```
<msubsup>                        <msup>
  <mi> x </mi>                     <msub>
                                     <mi> x </mi>
  <mn> 1 </mn>         →             <mn> 1 </mn>
                                   </msub>
  <mn> 2 </mn>                     <mn> 2 </mn>
</msubsup>                       </msup>
```

<mrow> Minimizing

```
<msqrt>                          <msqrt>
  <mrow>
    <mo> - </mo>       →           <mo> - </mo>
    <mn> 1 </mn>                   <mn> 1 </mn>
  </mrow>
</msqrt>                         </msqrt>
```

Applying Functions

```
<mi> f </mi>                     <mi> f </mi>
<mo> &#x2061; </mo>              <mrow>
<mrow>                             <mo> ( </mo>
  <mo> ( </mo>         →           <mi> x </mi>
  <mi> x </mi>                     <mo> ) </mo>
  <mo> ) </mo>                   </mrow>
</mrow>

                                  <mi>sin</mi>
<mi> sin </mi>                    <mrow>
                                    <mo>(</mo>
<mo> &#x2061; </mo>    →            <mi>x</mi>
                                    <mo>)</mo>
<mi> x </mi>                      </mrow>
```

Unifying Fences

```
                                 <mrow>
                                   <mo> [ </mo>
<mfenced open="[">                 <mrow>
                                     <mi> x </mi>
  <mi> x </mi>                       <mo> , </mo>
                       →             <mi> y </mi>
  <mi> y </mi>                     <mrow>
                                   <mo> ) </mo>
</mfenced>                       </mrow>
```

Indexing and Searching

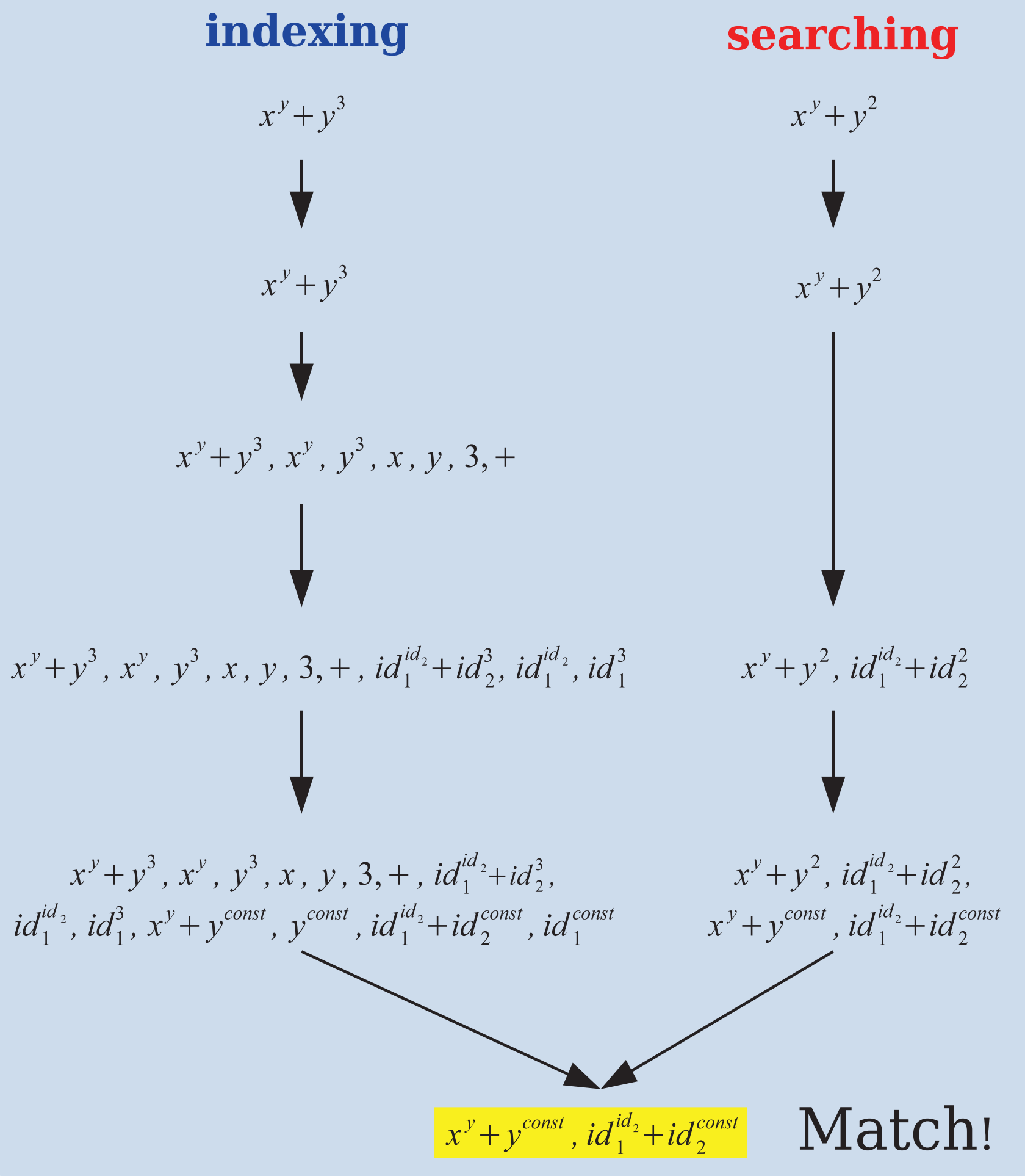
Our approach to searching mathematical content in documents is based on similarities of math structures through conventional full-text searching. As mathematical notation, e.g. expressions and formulae, is highly structured, we preprocess mathematical content in order to be processable by full-text searching methods. The preprocessing procedures include canoncalization, 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. Assigned weights affect the ordering of retrieved results. The factors that influence resulting weights of indexed subformulae are adjustable. The current setup respects our generic view of distance of extracted subformulae to their original trees. Different setup might influence the order of retrieved results significantly. There is no ideal set of factors, however, we want to reach to the optimal setup by repetitive evaluation.

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. 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.

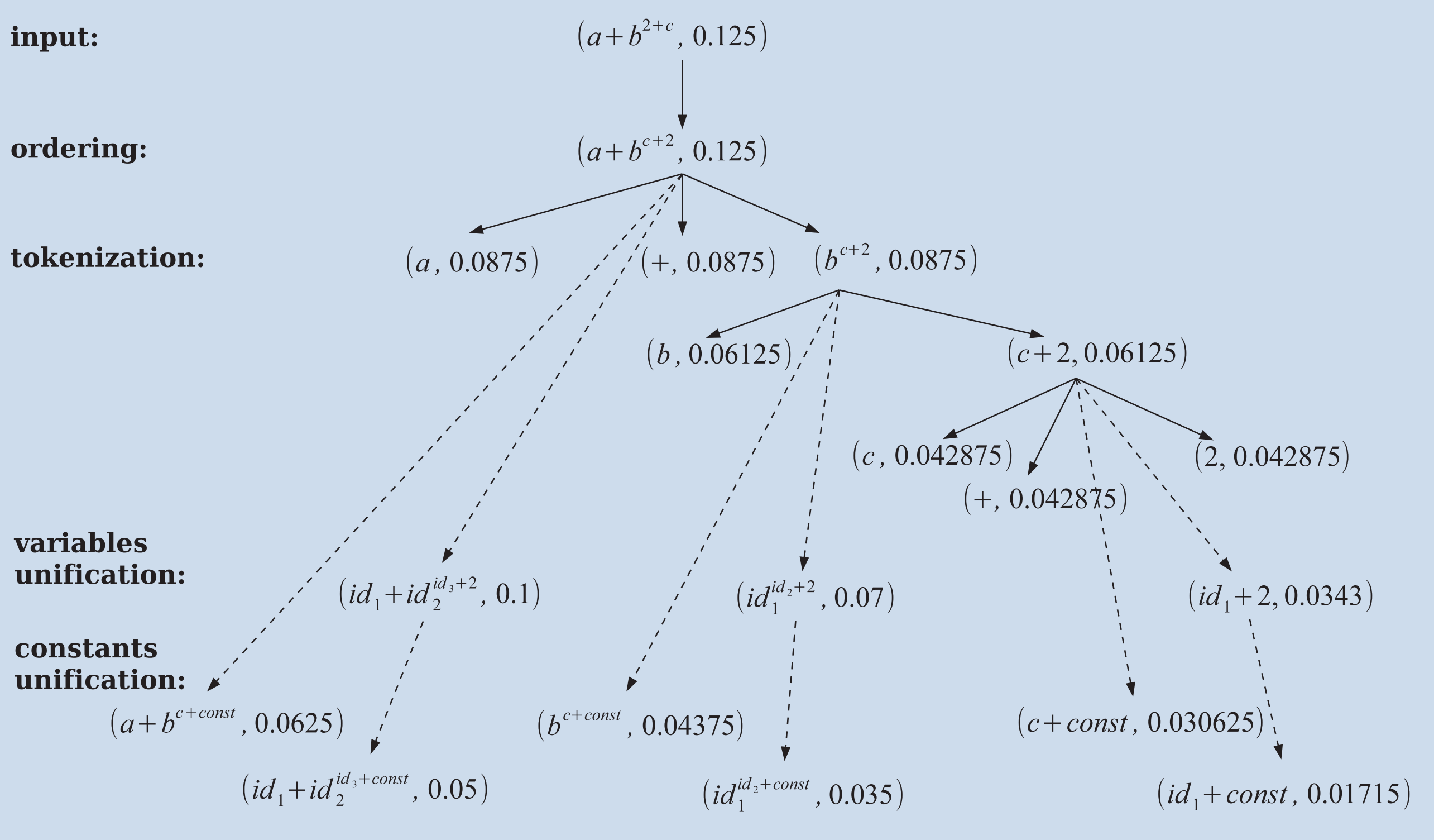
The very straightforward query interface of MIaS consists of only one input field. Users can type in textual queries together with math queries encoded using LaTeX notation as well as MathML notation. Query is on-the-fly visualized as `typeset' formula in user's web browser to allow users to verify the 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.

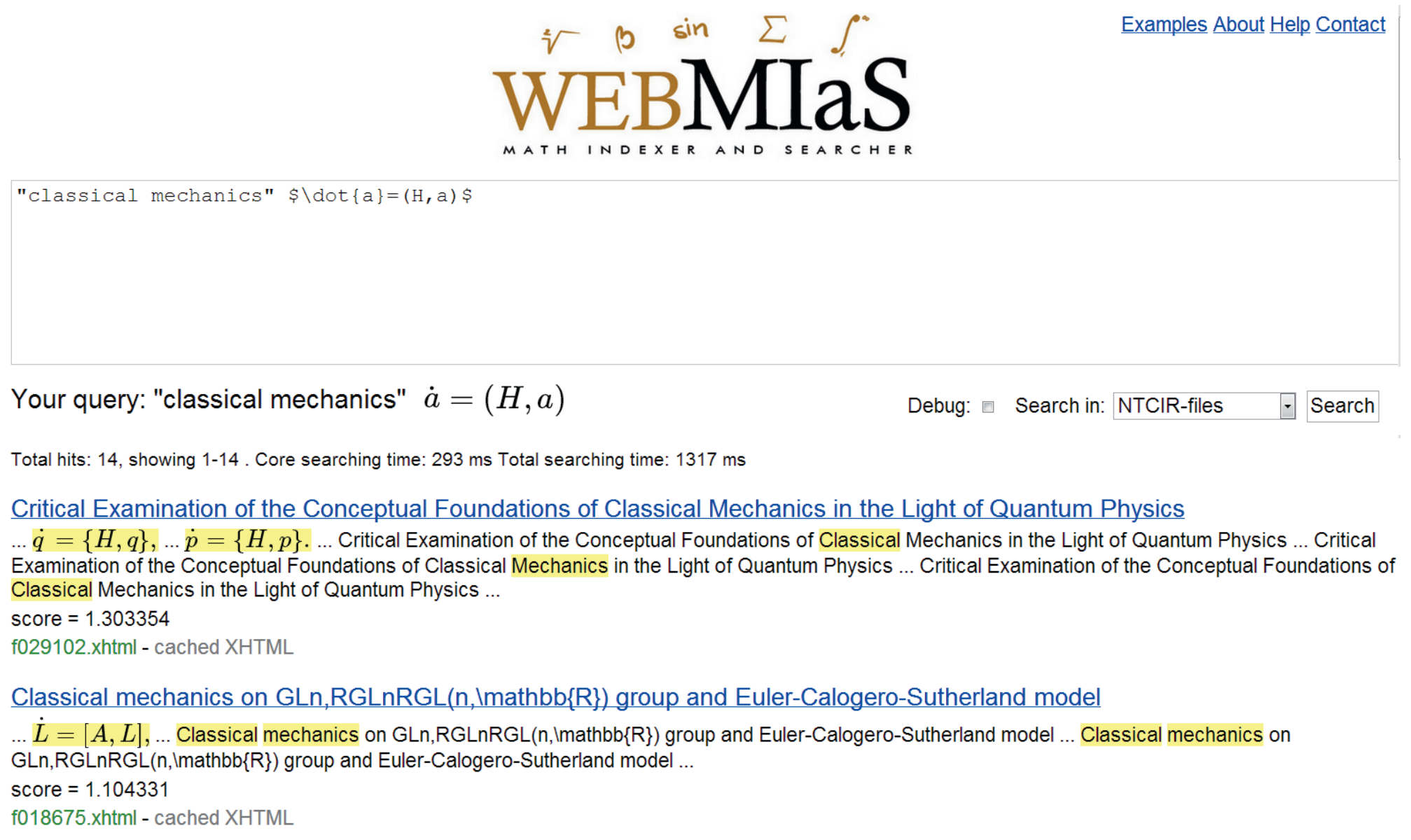
Alongside interactive web querying interface MIaS offers searching using web services. This is a indispensable feature for automated querying that was used to retrieve evaluation results for the NTCIR Math Task.

Workflow example


indexing
searching
x^y+y^3
x^y+y^2
x^y+y^3
x^y+y^2
$x^y+y^3,\ x^y,\ y^3,\ x,\ y,\ 3, +$
$x^y+y^3,\ x^y,\ y^3,\ x,\ y,\ 3, +,\ id_1^{id_2}+id_2^3,\ id_1^{id_2},\ id_1^3$
$x^y+y^2,\ id_1^{id_2}+id_2^2$
$x^y+y^3,\ x^y,\ y^3,\ x,\ y,\ 3, +,\ id_1^{id_2}+id_2^3,\ id_1^{id_2},\ id_1^3,\ x^y+y^{const},\ y^{const},\ id_1^{id_2}+id_2^{const},\ id_1^{const}$
$x^y+y^2,\ id_1^{id_2}+id_2^2,\ x^y+y^{const},\ id_1^{id_2}+id_2^{const}$
$x^y+y^{const},\ id_1^{id_2}+id_2^{const}$
Match!

Math Processing





"classical mechanics" $\dot{a}=(H,a)$

Your query: "classical mechanics" $\dot{a} = (H,a)$

Total hits: 14, showing 1-14 . Core searching time: 293 ms Total searching time: 1317 ms

Critical Examination of the Conceptual Foundations of Classical Mechanics in the Light of Quantum Physics
... $\dot{q} = \{H,q\}$, ... $\dot{p} = \{H,p\}$. ... Critical Examination of the Conceptual Foundations of Classical Mechanics in the Light of Quantum Physics ... Critical Examination of the Conceptual Foundations of Classical Mechanics in the Light of Quantum Physics ... Critical Examination of the Conceptual Foundations of Classical Mechanics in the Light of Quantum Physics ...
score = 1.303354
f029102.xhtml - cached XHTML

Classical mechanics on GLn,RGLnRGL(n,\mathbb{R}) group and Euler-Calogero-Sutherland model
... $\dot{L} = [A,L]$, ... Classical mechanics on GLn,RGLnRGL(n,\mathbb{R}) group and Euler-Calogero-Sutherland model ... Classical mechanics on GLn,RGLnRGL(n,\mathbb{R}) group and Euler-Calogero-Sutherland model ...
score = 1.104331
f018675.xhtml - cached XHTML

Results

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

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

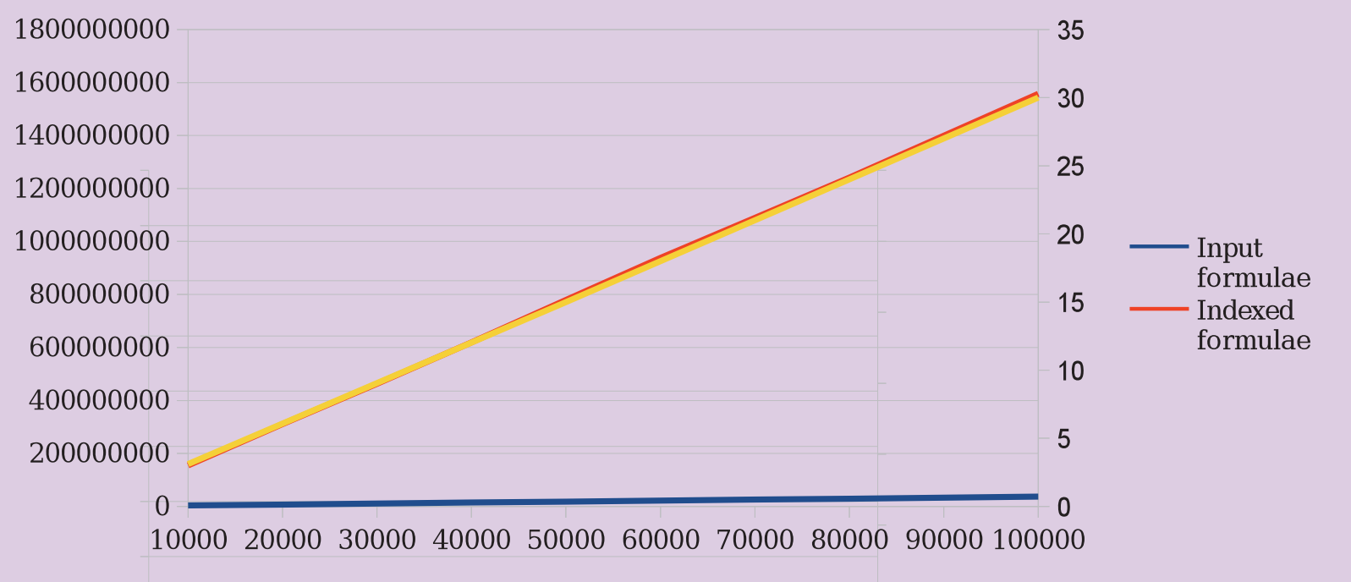
Run #	Query language
1	Presentation MathML
2	Content MathML
3	Presentation and Content MathML
4	TEX

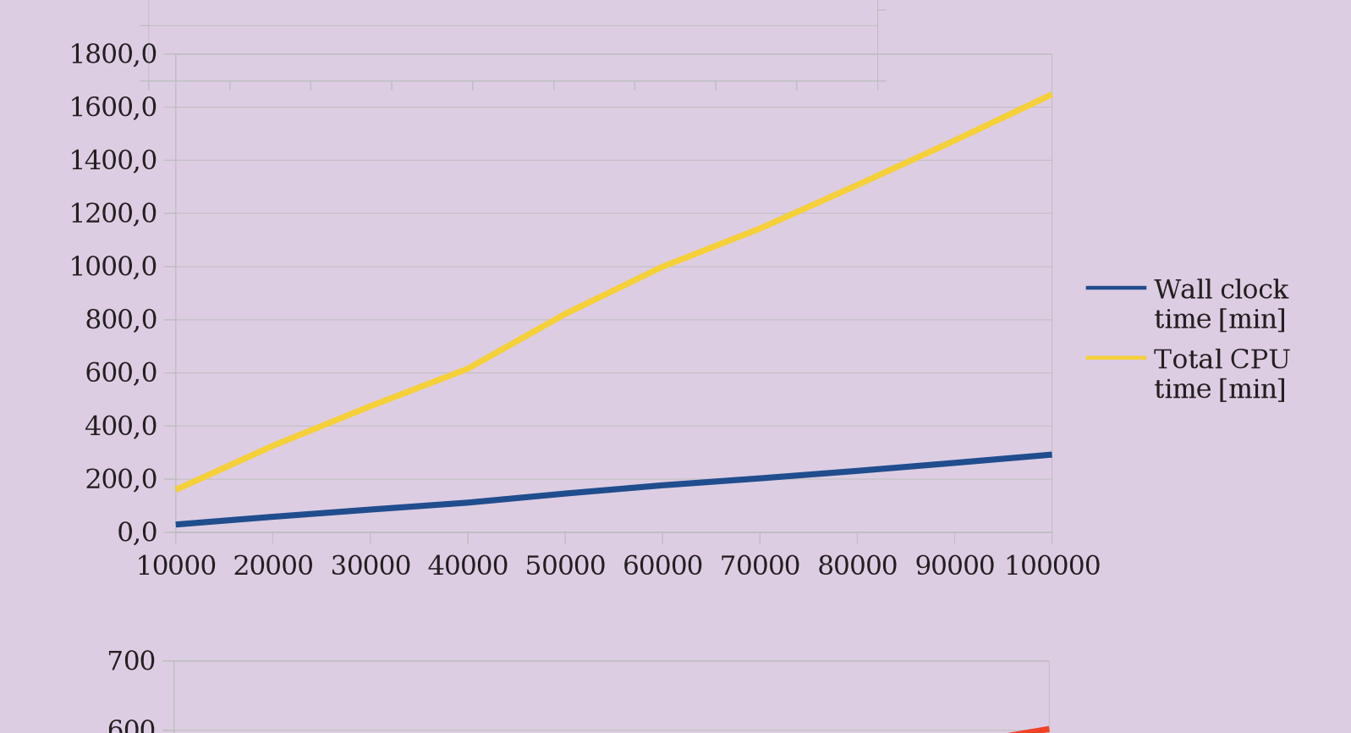
Table 3: Result metrics for submitted runs in Formula Search with Relevance Level ≥ 3 (Relevant)

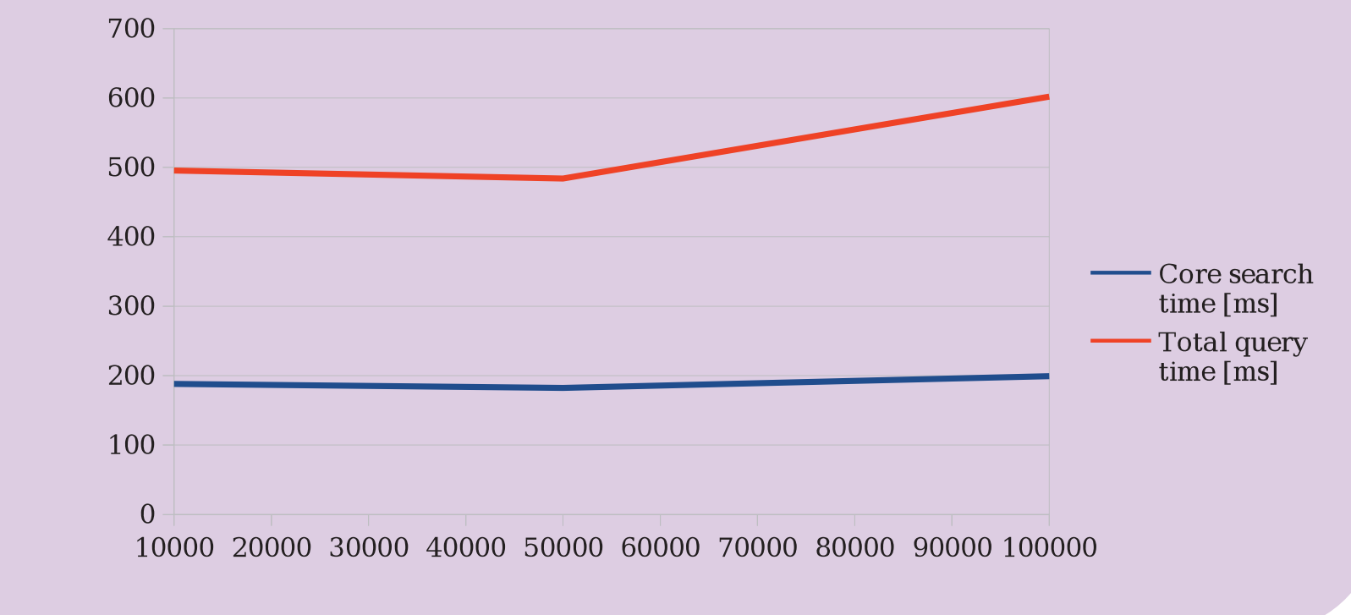
Metric	Run 1	Run 2	Run 4
P-10 avg	0.105	0.191	**0.219**
P-5 avg	0.133	0.229	**0.276**
MAP avg	0.060	0.112	**0.127**
Precision	0.109 (64/589)	**0.185** (92/496)	0.123 (96/778)

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







Evaluation

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 for the NTCIR-10. Full Text Search simulated the standard use of a search system – queries comprised of 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 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.

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 LaTeXML 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.

Overall scores of MIaS were above average of the Math Task results. 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'.

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. Our future goal is to create our own, gold standard evaluation collection. We find it a prerequisite to the further development of retrieval techniques.

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.

LÍŠKA, Martin and Petr SOJKA and Michal RŮŽIČKA. Similarity Search for Mathematics: Masaryk University team at the NTCIR-10 Math Task. In Proceedings of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Math Pilot Task

Acknowledgements: This work was partially supported by the European Union through its Competitiveness and Innovation Programme (Information and Communication Technologies Policy Support Programme, ‘Open access to scientific information’, Grant Agreement No. 250503, a project of the European Digital