Martin Líška; Petr Sojka; Michal Růžička; Petr Mravec

Web Interface and Collection for Mathematical Retrieval : WebMIaS and MREC

In: Petr Sojka and Thierry Bouche (eds.): Towards a Digital Mathematics Library. Bertinoro, Italy, July 20-21st, 2011. Masaryk University Press, Brno, Czech Republic, 2011. pp. 77--84.

Persistent URL: `http://dml.cz/dmlcz/702604`

Web Interface and Collection for Mathematical Retrieval

WebMIaS and MREC

Martin Líška, Petr Sojka, Michal Růžička, and Petr Mravec

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

Abstract. We demonstrate searching of mathematical expressions in technical digital libraries on a *MREC collection* of 439,423 real scientific documents with more than 158 million mathematical formulae. Our solution — *the WebMIaS system* — allows the retrieval of mathematical expressions written in TeX or MathML. TeX queries are converted on-the-fly into tree representations of Presentation MathML, which is used for indexing. WebMIaS allows complex queries composed of plain text and mathematical formulae, using MIaS (Math Indexer and Searcher), a math aware search engine based on the state-of-the-art system Lucene. MIaS implements proximity math indexing with a subformulae similarity search.

Keywords: math indexing and retrieval, mathematical digital libraries, information systems, information retrieval, mathematical content search, document ranking of mathematical papers, math text mining, WebMIaS, MIaS, Tralics, TeX, UMCL, Lucene

1 Introduction

The gateway to the vast treasures held in digital libraries' content is entered by *searching*. The Google generation is starting to demand a simple Google-like interface to access digital content, even on a global scale. The mainstream technologies and interfaces are developed only for plain text without support for mathematical formulae handling — documents are represented in a bag of words representation, in a simple vector space model.

Scientific and technical documents are full of indexes, exponents, and complex mathematical expressions, even in paper basic metadata, titles and abstracts. Our experience with Google Scholar shows that not handling mathematical expressions in citations causes severe problems. For example the paper by Kováčik and Rákosník [3] appears as more than twenty different papers there[1] mainly because of different and wrong (by different OCR) representation of mathematics in the paper metadata (title).

Although there have been several attempts to solve the mathematics search problem, none of them have, as yet, fulfilled the expectations. For example, Springer offers LaTeXSearch[2] based just on TeX math string matching,

[1] cf. http://scholar.google.com/scholar?q=Kovacik+Rakosnik   [2] http://www.latexsearch.com/

Petr Sojka, Thierry Bouche (editors): DML 2011, Towards a Digital Mathematics Library, pp. 77–84.
 ISBN 978-80-210-5542-1

Table 1: Documents collected from arXMLiv

arXMLiv transformation result class	**Quantity**
successful (no problem)	65,874
successful (warning)	291,879
complete with errors (missing macros)	81,670
All documents	439,423

which does not take into account the structural or semantical similarity of mathematical expressions at all.

We have created the web interface WebMIaS for our MIaS (Math Indexer and Searcher) system [6] indexing hundreds of thousands[3] of mathematical documents. We demonstrate a solution built on the state-of-the-art fulltext indexing engine Lucene ™ — we have added 'math-awareness' to it as a plug-in.

To test the system, we have created (Section 2) and indexed (Section 3 on the next page) the MREC collection of hundreds of thousands mathematical documents. In Section 4 on the facing page we describe necessary transformations needed during querying and indexing (canonicalization of MathML). The WebMIaS web interface is then presented in Section 5 on page 80. The reader finds final remarks in Section 6 on page 83.

2 Mathematical Retrieval Collection MREC

To evaluate our system, we have built a corpus of mathematical texts, called MREC. We downloaded documents from arXMLiv[4] [8], where TEX documents from arXiv.org are transformed into XML documents. For the representation of mathematical formulae, MathML, a W3C standard, is used. The documents used come from different scientific areas (Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics).

ArXMLiv[5] sorts transformed documents into several classes, based on the return value of transformation to MathML: successful, complete with errors, incomplete and none. MREC does not contain full arXiv, only documents from conversion classes successful and complete with errors (missing macros) — see Table 1. We have collected 439,423 documents in well-formed XHTML, containing mathematical formulae in valid MathML. We hope that this corpus might be used for benchmarking mathematical retrieval, thus we have named it MREC (Mathematical REtrieval Collection) and made it available for this purpose at [4].

In our web interface for math searches we currently use this corpus of real mathematical papers.

[3] LaTeXSearch currently searches only three million formulae. [4] `http://kwarc.info/projects/arXMLiv/`

[5] `http://arxmliv.kwarc.info/`

3 Math-aware Indexing

We have developed a math aware, full-text based search engine called MIaS (Math Indexer and Searcher). [6] It processes documents containing mathematical notation in Presentation MathML format, however, it filters out all unnecessary presentational elements as well as any other MathML notation (Content MathML or other markup). MIaS allows users to search for mathematical formulae as well as the textual content of documents.

Since mathematical expressions are highly structured and have no canonical form, our system pre-processes formulae in several steps to facilitate a greater possibility of matching two equal expressions with different notation and/or non-equal, but similar formulae. With an analogy to natural language searching, MIaS searches not only for whole sentences (whole formulae), but also for single words and phrases (subformulae down to single variables, symbols, constants, etc.). For every formula and its subformulae on the input, MIaS creates several differently generalized representations to allow similarity searching of mathematics. For calculating the relevance of matched expressions to the user's query, MIaS uses a heuristic weighting of indexed terms, which accordingly affects scores of matched documents and thus the order of results. Weights are assigned to the formula according to the complexity of the formula, its level in the input formula tree and level of generalization.

At the end of all these processing methods, formulae are converted from XML nodes to a compacted linear string form which can be handled by the indexing core.

4 System Workflow

The top-level indexing scheme is shown in Figure 1 on page 81. Document and query processing is done separately for plain text terms and mathematical terms. Indexing of mathematics is done by our Presentation MathML tokenizer implemented in Java for Apache Lucene™3.1, and Lucene Solr™ 3.1 taking advantage of open Lucene architecture.

MathML notation in the query and indexed documents is normalized into Canonical MathML [1] to increase precision of the system. For conversion into this normalized MathML format we are using the software library UMCL (Universal Maths Conversion Library). The main purpose of the UMCL toolset is the transcription of the MathML formulae to Braille national codes. Related to our task is also the need for MathML formulae unification. UMCL transformation of the MathML to Canonical MathML is carried out using a set of XSL stylesheets. This transformation was integrated into the WebMIaS system with only the slightest modifications — the UMCL transformation adds attributes in the form of `id="formula:xx"` to every node of the output MathML. This is not necessary for the WebMIaS purposes as it adds additional 'noise' to the formulae and increased size of the index. Thus, these attributes are not added to the Canonical MathML used by WebMIaS.

Our latest experiments with canonical forms of MathML generated by the UMCL transformation show that it not only increases fairness of similarity ranking, but also helps to match a query against the indexed form of MathML. For example, if the user asked the system for the

$$x^2 + y^2$$

formula using MathML of the form

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
    <msup>
        <mi>x</mi>
        <mn>2</mn>
    </msup>
    <mo>+</mo>
    <msup>
        <mi>y</mi>
        <mn>2</mn>
    </msup>
</math>
```

the system would not be able to find any similar formulae due to omission of the <mrow> element in the MathML. Provided that the MathML canonicalization of the query is done prior to the search, the canonical form of the query

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mrow>
    <msup>
        <mi>x</mi><mn>2</mn></msup>
    <mo>+</mo>
    <msup>
        <mi>y</mi><mn>2</mn></msup>
  </mrow>
</math>
```

results in 36,817 hits in MREC 2011.4.

For a user-friendly math-aware information retrieval demonstration, we have built web interface *WebMIaS* (see Figure 2 on page 82).

5 WebMIaS

WebMIaS demonstrates the possibility of querying mathematical content on a large-scale. This has been facilitated by the full indexation of the mathematical corpus MREC. In the user interface (UI) we tried to mimic the simplicity of Google. In addition to the standard textual query terms, mathematics terms (mterms) may appear in the query as well, adding to the document score with the weight depending on the similarity of matched formula to the queried one. Mterm could be either in MathML, or in TEX notation enclosed in two dollar signs. Since most mathematicians are used to using TEX compact

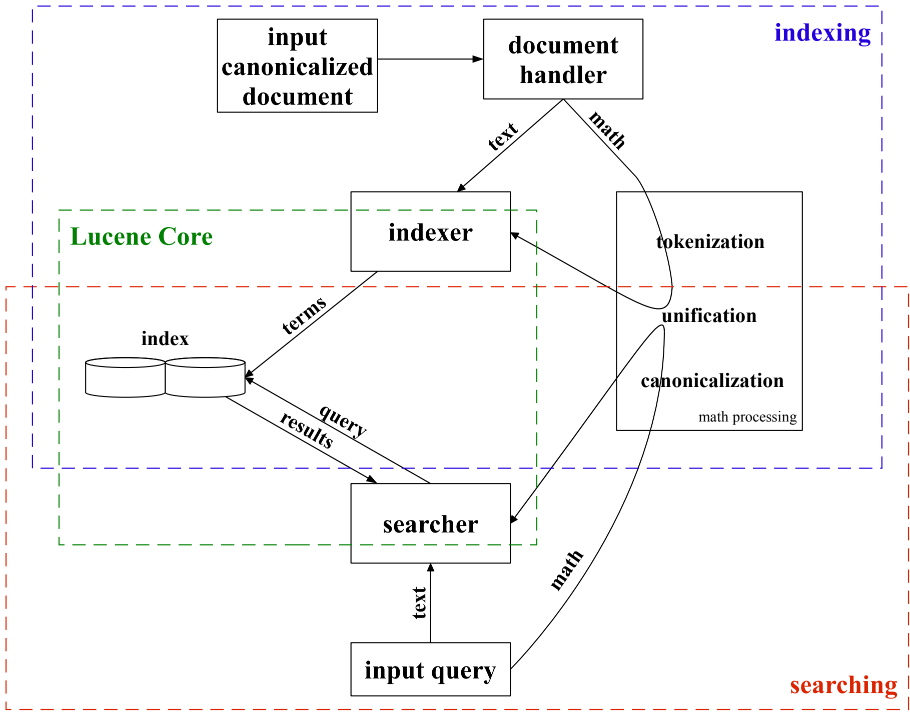

Fig. 1: Scheme of the system workflow

notation for mathematical formulae, we have implemented on-the-fly TEX to MathML conversion [7] of queries using Tralics [2] as a library. Furthermore, canonicalization of the both MathML and TEX input queries has been employed to improve querying and to avoid notation flaws restraining proper results retrieval. For the best visual experience of the search results, we incorporated a much requested snippet retrieval and mathematical match highlighting in the hit list for each matched document. This will also help us to evaluate the search results and to be able to tweak the whole indexing and searching process for better results. We additionally incorporated MathJax in the UI for a better rendering and look of displayed mathematical snippets, which will in turn enhance web browser support, since not all of the web browsers have natural MathML rendering capabilities.

As is shown in Table 2 on page 83, the performance of the system scales linearly. This gives feasible response times even for our billions of indexed subformulae. One has to be patient for small formulae, as they score/match in most documents. We also tried to measure the average query time of WebMIaS working over the MREC 2011.4 corpus. We queried the created index with a set

Fig. 2: WebMIaS web interface

of differently complex queries (mixed, non-mixed, more/less complex single/multiple formulae). The resulting average query time was 469 ms.

It is very difficult to evaluate the mathematical search result and verify the soundness of our design. For a given set of queries, there should exist beforehand a complete list of the documents ordered by their relevance to the query with which the actual results can be compared with.

We have applied an empirical approach to the evaluation so far using our WebMIaS demo interface which is publicly available at `http://nlp.fi.muni.cz/projekty/eudml/mias/`. It currently works on our mathematical corpus MREC version 2011.4 with 158,106,118 input formulae, 2,910,314,146 indexed (sub)formulae.

Table 2: Scalability test results (run on 448 GiB RAM, eight 8-core 64bit processors Intel Xeon™ X7560 2.26 GHz driven machine).

# Docs	Input formulae	Indexed formulae	Indexing run-time [ms]	Indexing CPU time [ms]
10,000	3,406,068	64,008,762	2,145,063	2,102,770
50,000	18,037,842	333,716,261	11,382,709	10,871,500
100,000	36,328,126	670,335,243	23,066,679	21,992,100
200,000	72,030,095	1,326,514,082	46,143,472	44,006,180
300,000	108,786,856	2,005,488,153	71,865,018	66,998,550
350,000	125,974,221	2,318,482,748	83,199,724	77,886,160
439,423	158,106,118	2,910,314,146	104,829,757	97,393,301

6 Conclusion

We have demonstrated the fully functioning information retrieval interface, WebMIaS, capable of retrieving both text and math from fulltexts in Presentation MathML. The system scales well and has got the power to be used in several digital libraries.

As our developments were motivated by future deployment in the EuDML[6] project [9], experience with WebMIaS results will be projected and employed in the EuDML UI. Another area of long-term research planned is supporting Content MathML, in a way similar to the current handling of Presentation MathML. The architectural design is suited to it, but as most of the math within EuDML will be in Presentation MathML taken from PDFs, this is not currently a high priority.

Acknowledgements. This work has been in part financed by the European Union through its Competitiveness and Innovation Programme (Information and Communications Technologies Policy Support Programme, "Open access to scientific information", Grant Agreement No. 250503).

References

1. Archambault, D., Moço, V.: Canonical MathML to Simplify Conversion of MathML to Braille Mathematical Notations. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (eds.) Computers Helping People with Special Needs, Lecture Notes in Computer Science, vol. 4061, pp. 1191–1198. Springer Berlin / Heidelberg (2006), `http://dx.doi.org/10.1007/11788713_172`
2. Grimm, J.: Producing MathML with Tralics. In: Sojka [5], pp. 105–117, `http://dml.cz/dmlcz/702579`
3. Kováčik, O., Rákosník, J.: On spaces $L^{p(x)}$ and $W^{k,p(x)}$. Czechoslovak Mathematical Journal 41, 592–618 (1991), `http://dml.cz/dmlcz/102493`
4. MREC — Mathematical REtrieval Collection, `http://nlp.fi.muni.cz/projekty/eudml/MREC/index.html`

[6] `http://eudml.eu`

5. Sojka, P. (ed.): Towards a Digital Mathematics Library. Masaryk University, Paris, France (Jul 2010), http://www.fi.muni.cz/~sojka/dml-2010-program.html
6. Sojka, P., Líška, M.: Indexing and Searching Mathematics in Digital Libraries – Architecture, Design and Scalability Issues. In: Davenport, J.H., Farmer, W., Rabe, F., Urban, J. (eds.) Proceedings of CICM Conference 2011 (Calculemus/MKM). Lecture Notes in Artificial Intelligence, LNAI, vol. 6824, pp. 228–243. Springer-Verlag, Berlin, Germany (Jul 2011)
7. Stamerjohanns, H., Ginev, D., David, C., Misev, D., Zamdzhiev, V., Kohlhase, M.: MathML-aware Article Conversion from LaTeX. In: Sojka, P. (ed.) Proceedings of DML 2009. pp. 109–120. Masaryk University, Grand Bend, Ontario, CA (Jul 2009), http://dml.cz/dmlcz/702561
8. Stamerjohanns, H., Kohlhase, M., Ginev, D., David, C., Miller, B.: Transforming Large Collections of Scientific Publications to XML. Mathematics in Computer Science 3, 299–307 (2010), http://dx.doi.org/10.1007/s11786-010-0024-7
9. Sylwestrzak, W., Borbinha, J., Bouche, T., Nowiński, A., Sojka, P.: EuDML—Towards the European Digital Mathematics Library. In: Sojka [5], pp. 11–24, http://dml.cz/dmlcz/702569