Presenting and Searching Mathematics in Digital Repositories

Peter Stanchev[1,2], Jiří Rákosník[3], Radoslav Pavlov[1], Georgi Simeonov[1]

[1] Institute of Mathematics and Informatics, BAS, Bulgaria
[2] Kettering University, Flint, USA
[3] Institute of Mathematics, CAS, Czech Republic
pstanche@kettering.edu, rakosnik@math.cas.cz,
radko@cc.bas.bg, gsimeonov@math.bas.bg

Abstract. The paper presents an overview of the current development of tools for search for mathematical formulae and their implementation in Digital Mathematical Libraries and reference databases such as zbMATH, MathSciNet and EuDML for mathematical scholarly literature.

Keywords: Digital Mathematics Library, Formula search

1 Introduction

In the light of the ever faster growing volume of scholarly literature in mathematics, web-based access to digital resources becomes absolutely essential for the work of mathematicians and other users of mathematical knowledge. These resources include digital libraries, reference databases and specialized web sites. An outline of recent development in Digital Mathematics Libraries is presented in [36].

While text search in digital resources is now commonplace, search for mathematical formulae, equations, theorems, and proofs is a challenge attracting attention of mathematicians, computer scientists and developers.

For typesetting their texts, mathematicians today typically use the typographic system TEX [17] (with its variants LATEX [20], AMS- LATEX [1] etc.), which however is not suitable for presentation on the web and involves too much ambiguity for the purpose of retrieval. MathML [27], OpenMath [6, 35], OMDoc/MMT [15, 34] and Mizar [30, 16] represent the main languages and systems for encoding and presenting mathematics for the web. They are employed by various special search engines for mathematical text, e.g., EgoMath [9], LATEXSearch [22], ActiveMath [23], MathWebSearch [19, 13, 18] or MIaS [37].

Although still essentially under development, these systems and engines are being implemented in digital mathematics libraries, reference databases and specialized math-oriented web services.

We shall give an overview of some of the systems mentioned above and tools with particular regard to their implementation in digital resources for information in mathematics.

2 Presenting and searching mathematical formula

TEX [17], one of the most sophisticated digital typographical systems in the world (with its variants LATEX [20], AMS-LATEX [1] etc.), was designed to allow everybody to produce high-quality prints, particularly of mathematical formulae, giving exactly the same results on all computers, at any point in time. However, while the possibility to write a formula in a non-unique way facilitates typesetting, it complicates presentation on the web and in fact does not allow detecting the semantics without the context.

Much effort has been put into overcoming this by constructing specialized languages and systems. The mathematical markup language MathML [27], an application of XML for describing mathematical notation and capturing both its structure and content, is now widely used for presentation of mathematical formulae on the web. The system OpenMath [6, 35] was designed to represent the semantics of mathematical objects. Both standards can be considered complementary in the sense that MathML provides a presentation format for mathematical objects, while OpenMath provides a mechanism for describing the semantics of mathematical symbols.

Mizar [30, 16] is a representation format for mathematics and a formal system for completing and verifying proofs written in the Mizar language. It operates on the Mizar Mathematical Library [30], which is one of the largest libraries of formalized and mechanically verified mathematics. It is harder to specify for machine manipulation. There is a translation of the Mizar library into the Open Mathematical Documents format [16], a semantic markup format for mathematical documents based on XML representation allowing to write down the meaning of texts about mathematics.

The search engines EgoMath [9] and LATEXSearch [22] can search for mathematical formulae written in LaTeX and simple text providing a list of matched documents and snippets with highlighted matched terms. The search back-end of the former one is based on the Apache Solr search platform while the indexer and the front-end are standalone applications.

In the Digital Library of Mathematical Functions [8, 29], an online project at the National Institute of Standards and Technology for developing a major resource of mathematical reference data for special functions and their applications, mathematical formulae are converted to text and indexed. The search string is similar to LATEX commands and is converted to string before searching. The modified representation corresponds to the content markup in the sense of MathML or OpenMath. A similar approach is applied in the intelligent tutoring system ActiveMath [23].

The LaTeXML system [21] is being developed in the Digital Library of Mathematical Functions project to transform LaTeX sources into content and presentation MathML. In [12] an extension, the LaTeXML Daemon allowing efficient, scalable and on-the-fly processing is presented. Its strength was demonstrated by converting 1.5 million abstracts from the zbMATH database [39].

MathWebSearch [19, 13, 18] is a content-based full-text search engine that indexes MathML formulae, using a technique derived from automated theorem proving. It combines exact formula matching with the full text search capabilities of ElasticSearch [10] for searching in mathematical texts. MathWebSearch consists of three system components: a set of web crawlers that periodically scan the web, identify, and download suitable mathematical data, a search server encapsulating the index, and a

web server communicating the results to the user. The search engine can be employed by various front-ends like TeMaSearch (combined text and math search for the zbMATH database), zbMATH Search (content-oriented search engine for the formulae in the 3.3 million reviews and abstracts in zbMATH), XLSearch (search engine for spreadsheet formulae), SentidoSearch (multi-format input front-end for search queries based on the Sentido system).

MaTeSearch [2] is a new type of search engine, which can handle a mathematical query and a text query at the same time. It is built on top of MathWebSearch and of the text-based search engine Nutch [4] built on the open-source Lucene architecture [3]. MaTeSearch connects to these components in order to obtain two different sets of results which are merged by intersection.

The Math Search Engine [14], like MathWebSearch, is using the structure of mathematical formulae. It makes inverted indexes by using the formula described in MathML language or in OpenMath.

WebMIaS [24] allows the retrieval of mathematical expressions written in TEX or MathML converting TeX queries on-the-fly into tree representations of presentation MathML, which is used for indexing. The queries can be composed of plain text and mathematical formulae. WebMIaS uses the math aware search engine MIaS [37] based on the state-of-the-art system Lucene. MIaS implements proximity math indexing with a subformula similarity search.

3 Implementation in Mathematical Text Repositories

3.1 zbMATH Database

Zentralblatt MATH (zbMATH) [39] is the world's most comprehensive and longest-running abstracting and reviewing service in pure and applied mathematics. It provides easy access to bibliographic data, reviews and abstracts of more than 3 million mathematical scholarly texts published since 1868. The database zbMATH contains about 1,650,000 direct links to electronic versions of the indexed publications, to the publishers' websites and/or to electronic libraries with open access to the full texts.

The structured search in zbMATH allows for free combination of different query types including the formula search based on the MathWebSearch system. MathML is a recommendation of the W3C math working group and part of HTML5. The retrieval system is now based on HTML5, allowing a quick and correct display of mathematical formulae in MathML for almost all modern web browsers. For older browsers MathJax [26] is used, a cross-browser JavaScript library released before HTML5, that displays mathematical notation in web browsers using MathML and LaTeX.

3.2 Mathematical Reviews and MathSciNet

Mathematical Reviews [25] is a reference journal published by the American Mathematical Society since 1940. Like zbMATH, it contains bibliografic data and brief synopses of scholarly mathematical publications. MathSciNet [28], the electronic version of Mathematical Reviews, presents a fully searchable database with many tools designed to help navigating the mathematical sciences literature, including: re-

views written by a community of experts, bibliographic listings dating back to the early 1800s, links to articles, journals, and publishers, linked reference lists, citation information on articles, books, and journals. MathSciNet also offers citation searching, by author or by journal. Search through mathematical formulae is not implemented.

MR Lookup provides bibliographic matching in the Mathematical Reviews database. It uses any of the following fields: ISSN, Journal, Author, Volume, Issue, Page, Year and/or Title. An interactive version [32] allows user to enter data in fields and receive up to three items that match their data. Users may add more information to their search if too many items are returned. If no items are returned they may reduce or modify the data.

Batch MR Lookup is intended for those who use a script to assemble bibliographic data for queries and process the results from MR Lookup. The Batch MR Lookup Query API includes 11 fields: ISSN, Journal title or abbreviation, Author name(s), Volume number, Issue ID number, Initial Page number, Year, Resource type, User supplied key, Identifier (Mathematical Reviews Number), and Item title.

MRef [31] is a tool for creating standard references with links to MathSciNet. The user enters the reference (often a portion of the reference is sufficient) for MRef to recognize the corresponding entry in MathSciNet.

3.3 EuDML – the European Digital Mathematics Library

The EuDML [11] is a system that provides a common framework, standards and services for unified seamless access to the distributed heterogeneous local digital repositories containing relevant mathematical literature published in Europe including periodicals, selected monographs and conference proceedings from the past as well as the currently produced mathematical publications. It was built in a project of 13 partners from 9 European countries partly supported by the EU [38]. Among the partners are the Bulgarian Digital Mathematical Library (BulDML) [5] and the Czech Digital Mathematical Library (DML-CZ) [7]. Both preserve national heritage in mathematics and offer free access to it,, provide their content to the EuDML and meet the requirements and standards for compliance and interoperability that have been developed in the project.

The numerous services provided to users by the EuDML comprise the structured search including formula search based on the MIaS system.

In EuDML a tool for working with formulae was created. A valid XML file that can contain mathematical formulae both as textual TEX formulae, within the text data of various elements, or formatted as elements with its internal EuDML DTD structure, containing a TEX-encoded version of the formula is analyzed. The result is a MathML representation of the formula. The tool written in Java identifies each formula in the input string, generates a standard NLM structure [33] for each of them, and returns a Java DOM Element containing the result.

LaTeX formulae in queries are converted on the fly to MathML and checked against a special index. Fuzzy matching and ranking is implemented in math search. Not only exact matches are found but also proximity subformulae matches are counted (with lower hit ranking).

The content in EuDML comes from different sources. Also the queries can be written in LaTeX as well as directly copy-pasted MathML. These different inputs produced by different converters and programmes can vary slightly in their notations. To make the searching more robust, math inputs from different sources are canonicalized to obtain one unified internal MathML representation. This representation does not suffer from MathML's ambiguities and contains only information necessary to the formula's meaning, which allows for higher probability of matching two equivalent formulae created in different ways.

4 Conclusions

The semantic search for mathematical formulae and other mathematical objects in digital documents represents a complex task which attracts the attention of mathematicians, computer scientists and developers. The recent progress made in this domain allows partial yet useful implementation of the corresponding tools in digital mathematics libraries and databases of mathematical knowledge. The fully satisfactory solution presents a challenge which will require much further research and concerted effort from scientists and developers.

References

1. AMS- LATEX http://www.ams.org/publications/authors/tex/amslatex/.
2. Anca, Ş.: MaTeSearch. A combined math and text search engine. Bachelor's thesis, Jacobs University Bremen, 2007. http://www.eecs.jacobs-university.de/archive/bsc-2007/anca.pdf.
3. Apache Lucene. https://lucene.apache.org/.
4. Apache Nutch. http://nutch.apache.org/.
5. Bulgarian Mathematics Digital Library, http://sci-gems.math.bas.bg/.
6. Buswell, S., Caprotti, O., Carlisle, D. P., Dewar, M. C., Gaetano, M., Kohlhase, M.: The Open Math standard, version 2.0. Technical report. The Open Math Society, 2004. http://www.openmath.org/standard/om20.
7. Czech Digital Mathematics Library, http://dml.cz/.
8. Digital Library of Mathematical Functions. http://dlmf.nist.gov/.
9. EgoMath. http://www.swmath.org/software/9766.
10. Elastic Search. https://www.elastic.co/.
11. EuDML. The European Digital mathematics Library. http://eudml.org/.
12. Ginev, D., Stamerjohanns, H., Miller, B. R., Kohlhase, M.: The LaTeXML Daemon: Editable Math on the Collaborative Web. In J. Davenport, W. Farmer, F. Rabe, J. Urban (eds.): Intelligent Computer Mathematics. 18th symposium, Calculemus 2011, and 10th international conference, MKM 2011, Bertinoro, Italy, July 18–23, 2011. Proceedings. Lecture Notes in Computer Science 6824. Lecture Notes in Artificial Intelligence, vol. 6824. Springer, 2011, pp. 292–294.
13. Hambasan, R., Kohlhase, M., Prodescu, C.: MathWebSearch at NTCIR-11. Proceedings of the 11th NTCIR Conference, December 9–12, 2014, Tokyo, Japan, pp. 1114–119.
14. Hijikata, Y., Hashimoto, H., Nishida, S.: Search mathematical formulas by mathematical formulas. In M. J. Smith, G. Salvendy (eds.): Symposium on Human Interface 2009, Held

as Part of HCI International 2009, San Diego, CA, USA, July 19–24, 2009, Proceedings, Part I. Lecture Notes in Computer Science, vol. 5617. Springer, 2009, pp. 404–411.
15. Iancu, M., Kohlhase, M., Prodescu, C.: Representing, archiving, and searching the space of mathematical knowledge. In H. Hong and C. Yap (eds.): Mathematical software – ICMS 2014. 4th international congress, Seoul, South Korea, August 5–9, 2014. Proceedings. Lecture Notes in Computer Science, vol. 8592. Springer, 2014, pp. 26–30.
16. Iancu, M., Kohlhase, M., Rabe, F., Urban, J.: The Mizar Mathematical Library in OMDoc: Translation and Applications. J. Autom Reasoning 50, no. 2 (2013), 191–202.
17. Knuth, D. E.: The TEXbook. Addison-Wesley Professional, 1984.
18. Kohlhase, M., Matican, B. A., Prodescu, C.-C.: MathWebSearch 0.5: Scaling an open formula search engine. In J. Jeuring, J. A. Campbell, J. Carette, G. Dos Reis, P. Sojka, M. Wenzel, V. Sorge (eds.): Intelligent computer mathematics. 11th international conference, AISC 2012, 19th symposium, Calculemus 2012, 5th international workshop, DML 2012, 11th international conference, MKM 2012, systems and projects, held as part of CICM 2012, Bremen, Germany, July 8–13, 2012. Proceedings. Lecture Notes in Computer Science, vol. 7362. Lecture Notes in Artificial Intelligence. Springer, 2012, pp. 342–357.
19. Kohlhase, M., Sucan, I.: A search engine for mathematical formulae. In T. Ida, J. Calmet, D. Wang (eds.): Artificial intelligence and symbolic computation. 8th international conference, AISC 2006, Beijing, China, September 20–22, 2006. Proceedings. Lecture Notes in Computer Science, vol. 4120. Lecture Notes in Artificial Intelligence, Springer, 2006, pp. 241–253.
20. Lamport, L.: LaTeX: A document preparation system. Addison-Wesley Professional, 1994.
21. LaTeXML. A LaTeX to XML/HTML/MathML Converter. http://dlmf.nist.gov/LaTeXML/.
22. LaTeXSearchBeta. http://latexsearch.com/.
23. Libbrecht, P., Melis, E.: Methods for access and retrieval of mathematical content in ActiveMath. In N. Takayama, A. Iglesias (eds.): Mathematical software – ICMS 2006. Second international congress on mathematical software, Castro Urdiales, Spain, September 1–3, 2006. Proceedings. Lecture Notes in Computer Science, vol. 4151. Springer, 2006, pp. 331–342.
24. Líška, M., Sojka, P., Růžička, M., Mravec, P.: Web Interface and Collection for Mathematical Retrieval: WebMIaS and MREC. In P. Sojka, T. Bouche (eds.): Towards a Digital Mathematics Library. Bertinoro, Italy, July 20–21st, 2011. Masaryk University Press, Brno, Czech Republic, 2011, pp. 77–84. http://dml.cz/dmlcz/702604.
25. Mathematical Reviews. http://www.ams.org/mr-database/.
26. MathJax. https://www.mathjax.org/.
27. MathML. W3C Math Home. http://www.w3.org/Math/.
28. MathSciNet. Mathematical Reviews. http://www.ams.org/mathscinet/.
29. Miller, B., Youssef, A.: Technical aspects of the digital library of mathematical functions. Annals of Mathematics and Artificial Intelligence, 38 (2003), no. 1–3, 121–136.
30. Mizar. http://mizar.org/.
31. MRef. http://www.ams.org/mathscinet-mref
32. MR Lookup. http://www.ams.org/mrlookup.
33. NLM Journal Archiving and Interchange Tag Suite. http://dtd.nlm.nih.gov/.
34. Open Mathematical Documents. https://trac.omdoc.org/OMDoc/.
35. OpenMath. http://www.openmath.org/.
36. Rákosník, J., Stanchev, P., Pavlov, R.: Recent Developments in Digital Mathematics Libraries. IR. Pavlov, P. Stanchev (eds.): Digital Presentation and Preservation of Cultural

and Scientific Heritage. Proceedings of the international conference, Veliko Tarnovo, Bulgaria, 18–21 September 2014. Vol. 4. Sofia, 2014, pp. 59–68.
37. Sojka, P., Líška, M.: Indexing and searching mathematics in digital libraries – architecture, design and scalability issues. In J. Davenport, W. Farmer, F. Rabe, J. Urban (eds.): Intelligent Computer Mathematics. 18th symposium, Calculemus 2011, and 10th international conference, MKM 2011, Bertinoro, Italy, July 18–23, 2011. Proceedings. Lecture Notes in Computer Science 6824. Lecture Notes in Artificial Intelligence, vol. 6824. Springer, 2011, pp. 228–243.
38. The Project of the European Digital Mathematics Library. http://project.eudml.org/.
39. zbMATH. https://zbmath.org/.