Why math search is more relevant now than ever?

Petr Sojka et al.

Masaryk University, Faculty of Informatics, Brno, Czech Republic
<sojka@fi.muni.cz>

Portmouth, UK

May 21st, 2012

Why Search?

I have a dream. [M.L. King, Jr., 1963]

Vast amounts of [moving] contents in digital libraries: from browsing to *search*; from static links to indirect search links.

Searching is crucial part of *accessibility* of the great ideas around, carved into 0s and 1s.

Why Math Search (MIR)?

A picture is worth thousands words.

The speeding up world is moving towards graphics (user interfaces and visualization).

"A math formulae is worth of hundreds of words." (Ross Moore)

There are papers with more formulae than plain text.

Why math search is more relevant *now* than ever? (cont.)

- Because of G? (G as in Google, Globalization,…).

- The *vast* treasure of mathematical papers; 140,000 new papers in Zentralblatt MATH expected this year. All mathematics ever publisher is estimated at 100,000,000 pages (3,500,000 articles).

- Search – crucial part; search is a *gate* to this knowledge; DML without math-aware search is an oxymoron.

- Text and keyword based search? Even picture search? No problem (Google, review databases); *success*.

- Mathematics formulae (structure) search? It *is* a problem (either in Google or in the review databases); more or less a *failure so far*.

Motivation for MSE (including formulae) – cont.

prof. James Davenport, CEIC member, MKM2011 PC chair, on panel at EuDML workshop in Bertinoro as a reply to the question "what functionality and incentives would made a working mathematician to login and use a modern DML as EuDML?":

"Math formulae search."

Why math *search* is more relevant now than ever?

- Allowing formulas in queries helps to *disambiguate and narrow* search.
  Sometimes the only difference among set of notions/key words would be
  in a math formula.

- Example 1: knowing the solution of partial differential equation in $L^1(\mathbb{C}^3)$,
  is there one in $L^2(\mathbb{C}^5)$?

- Example 2: historians may want to follow the history of a (class of) formula(s)
  across languages and vocabularies (e.g. same objects studied/used by
  physicists and mathematicians under different names).

- Imagine your favourite ebook math textbook being [TEX]-search aware—e.g.
  your search app supports math formulae search.

We do not start from scratch

Compare `google.com/search?q=Einstein` with math-aware search of `Einstein+$E=mc^2$` over arXiv.

Take-off message from this part of talk: *Yes, you can!*

I have a dream. [M.L. King, Jr.]

I hope

Yes, we can!

Towards math search engine (MSE) – existing players

- Niche market for big players (as Google), attempts to solve by publishers (LaTeXSearch by Springer).

- Many challenges: heterogenity of math representation, notation, semantics handling, no established and accepted user interface and query language.

- Numerous attempts to solve the problem: MathDex, EgoMath, LATEXSearch, LeActiveMath, DLMF equation search, MathWebSearch, but none accepted by the community as *the* MSE.

Existing systems – pros and cons

- **MathDex**: formerly MathFind * seven digit figure NSF grant by Design Science (Robert Miner) * Lucene based, indexing n-grams of presentation MathML * pioneering conversion effort

- **EgoMath and EgoMath2**: based on full text web search system Egothor * presentation MathML for indexing * idea of formulae augmentation, α-equivalence algorithms and relevance calculation

- **LaTeXSearch**: MSE offered by Springer * closed source * only for LaTeX math string approximate match based on strings * no formulae structure matching * small database: 3 million formulae from 'random' sources

- **LeActiveMath**: indexing string tokens from OMDoc with OpenMath semantic notation * *only* for documents authored for LeActiveMath learning environment

- **DLMF**: *only* for documents authored for DLMF in special markup * equation search

- **MathWeb Search**: semantic approach – uses substitution trees – not based on full text searching * supports Content MathML and OpenMath * problem with acquiring semantic data

MIaS — Math Indexer and Searcher

- math-aware, full-text based search engine
- joins textual and mathematical querying
- MathML *or* TEX input

How to write query

```
$x^2+y^2$ exponential distribution
```

Search in: MREC 2011.4.439 ▾   Search

Total hits: 15973, showing 1- 30. Searching time: 584 ms

Andreev bound states in normal and ferromagnet/high-Tcc superconducting tun ...
... close from the [110] surface when the symmetry is $d_{x^2+y^2}$.
score = 1.1615998
arxiv.org/abs/cond-mat/0305446 - cached XHTML

Particle trajectories and acceleration during 3D fan reconnection
... at $\sqrt{(x^2+y^2)}$ =1 and ...
score = 1.0577431
arxiv.org/abs/0811.1144 - cached XHTML

Pairing symmetry and long range pair potential in a weak coupling theory of ...
... does not mix with usual $s_{x^2+y^2}$ symmetry gap in an anisotropic band structure.
score = 1.0254444
arxiv.org/abs/cond-mat/9906142 - cached XHTML

Dual world of T_EX and MathML

Math for people: T_EX notation wins and is used by people (mostly AMSL^AT_EX fits most needs).

Math for software applications: MathML wins and is used by most computer algebra systems, browsers, in workflow of DTP systems…

Dual world of query language and indexing language

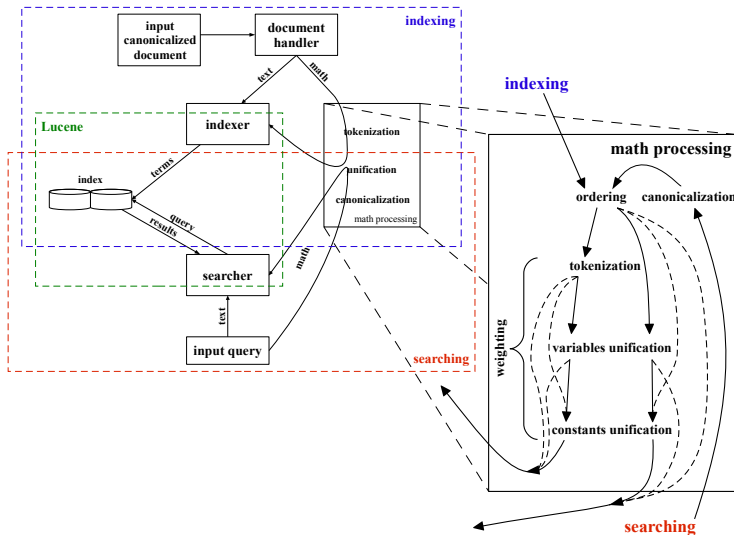In text retrieval: Indexing word stems only instead of world forms.

T$_E$Xbook's Concert invitation example: there is a name of Czech composer of a song in the index that even does not appear in the invitation.

From text to math: the same idea explored for math (e.g. having dozen of representations of a formula in the index).
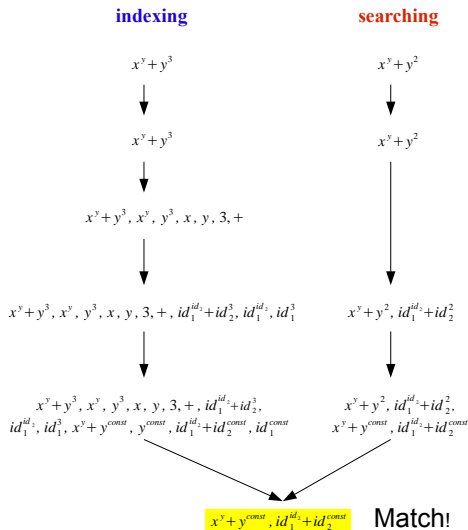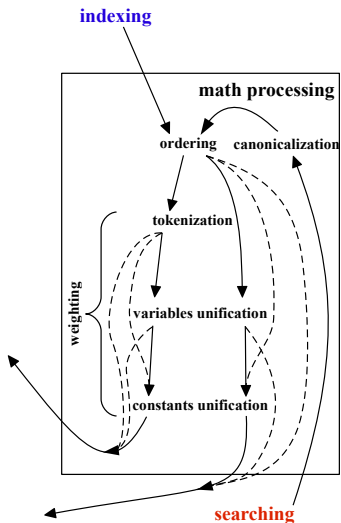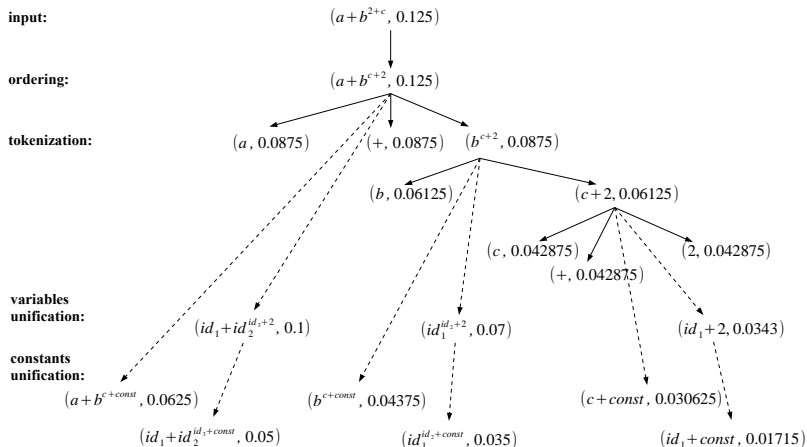
MSE overall design
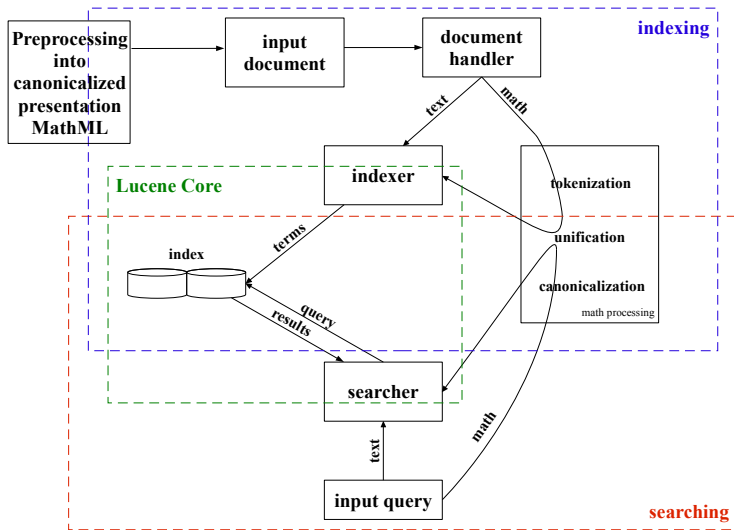
Math indexing design

Example

Formula processing example – subformulae weighting

input:
$$(a+b^{2+c},\ 0.125)$$

ordering:
$$(a+b^{c+2},\ 0.125)$$

tokenization:
$$(a,\ 0.0875) \quad (+,\ 0.0875) \quad (b^{c+2},\ 0.0875)$$

$$(b,\ 0.06125) \qquad (c+2,\ 0.06125)$$

$$(c,\ 0.042875) \qquad (2,\ 0.042875)$$

$$(+,\ 0.042875)$$

variables unification:
$$(id_1+id_2^{id_3+2},\ 0.1) \qquad (id_1^{id_2+2},\ 0.07) \qquad (id_1+2,\ 0.0343)$$

constants unification:
$$(a+b^{c+const},\ 0.0625)$$

$$(id_1+id_2^{id_3+const},\ 0.05) \qquad (b^{c+const},\ 0.04375) \qquad (c+const,\ 0.030625)$$

$$(id_1^{id_2+const},\ 0.035) \qquad (id_1+const,\ 0.01715)$$

Implementation

- Java

- Lucene 3.1.0

- Mathematical part implements Lucene's interface Tokenizer – able to integrate to any Lucene based system

- MIaS4Solr plugin was created for the use in Solr in EuDML

- Textual content – processed by StandardAnalyzer

Data used for evaluation: MREC corpus

- Mathematics REtrieval Corpus (MREC, version 2011.4.439)

  - 439,423 documents (originated from arXMLiv [8], validated, enriched with metadata for snippet generation)

  - Uncompressed size 124 GB, compressed 15 GB

  - 158 million input formulae, 2.9 billion subexpressions indexed (Lucene index size 47 GB)

- For more information see paper (DML 2011, Bertinoro) [10] and home page of MREC subproject http://nlp.fi.muni.cz/projekty/eudml/MREC/index.html

Scalability (tested on MREC 2011.4.439)

- Indexing time: 1,378.82 min (23 hours, down to 9 h with threads)

- Average query time: 469 ms

- Overall index size 47 GB (most of it math entries)

- Linear time scale – still seems feasible for a digital library

Search demonstration

EuDML

The EUROPEAN DIGITAL MATHEMATICS LIBRARY
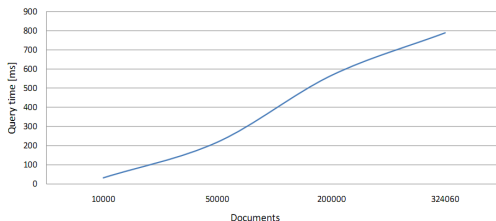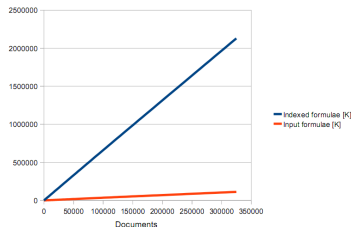
Help About

How to write query

```
<math><mrow><msup><mi>x</mi> <mn>2</mn> </msup><mo>+</mo><msup><mi>y</mi> <mn>2</mn> </msup></mrow></math>
```

Canonicalized MathML 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>
```

Search in: [MREC 2011.4.439 ▾] [Search]

Total hits: 36817, showing 1- 30. Searching time: 116 ms

Finite Precision Measurement Nullifies Euclid's Postulates
... and the unit circle $x^2 + y^2 = 1$ are both dense but they do not intersect, in contradiction to Euclid's postulates ...
score = 3.2980976
arxiv.org/abs/quant-ph/0310035 - cached XHTML

COMMENT ON RECENT TUNNELING MEASUREMENTS ON Bi22Sr22CaCu22O88
... gap, (b) s-wave gap, and (c) $s_{x^2+y^2}$ gap.
score = 1.6012012

Portsmouth University, May 21st, 2012: Why math search is more relevant now than ever

Formulae search demonstration comments

Demo web interface: http://aura.fi.muni.cz:8085/webmias/

- MathML/T$_{\text{E}}$X input (Tralics [2] for conversion to MathML [7])

- Canonicalization of the query – problems with UMCL library [1]

- Matched document snippet generation

- MathJax for nicer math rendering and better portability

MIaS already integrated in the the beta version of EuDML system.

Conclusions

- Scalable solution for math formulae search researched, implemented, tested and integrated into current version of EuDML system!

- MIaS project pages – http://nlp.fi.muni.cz/projekty/eudml/mias

Future work

- Preprocessing from TeX, PDF,...

- `copypaste` package (storing TeX math code into PDF as second layer with `/ActualText` (for indexing purposes): typesetters may use in their workflows.

- Improved MathML canonicalization and new preprocessing filters, test on more EuDML data.

- Weighting optimization (by machine learning).

- Query relaxation ("Did you mean…").

- Addition of Content MathML tree indexing?

- Mathematical equivalence computation via symbolic algebra system?

Summary

MIaS will hopefully become *the* MSE used by the community. Our hope is based on these features:

- *text+math IR compatible*, accepting both TEX and MathML formats (fits mathematician's needs)

- new math formulae similarity (weighting) approach compatible with *both presentation (structure) and content (semantic)* MathML

- *scalable* (index with almost 3 billion subformulae tested)

- *Lucene/Solr compatible* system employed and *used in EuDML will hit the masses* ;-).

For more information see papers in SpringerLink (MKM 2011, Bertinoro) [5] and ACM DL (DocEng 2011, Mountain View) [6].

Acknowledgments and questions?
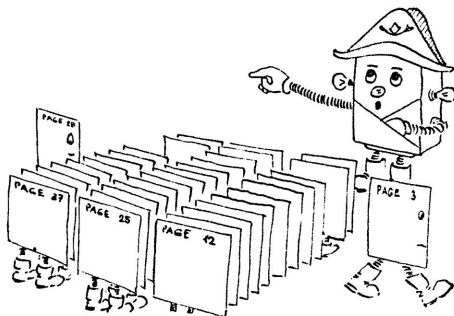
Archambault, D., Moço, V.: Canonical MathML to Simplify Conversion of MathML to Braille Mathematical Notations. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (eds.) Computers Helping People with Special Needs, Lecture Notes in Computer Science, vol. 4061, pp. 1191–1198. Springer Berlin / Heidelberg (2006), <http://dx.doi.org/10.1007/11788713_172>

Grimm, J.: Producing MathML with Tralics. In: Sojka [4], pp. 105–117, <http://dml.cz/dmlcz/702579>

MREC – Mathematical REtrieval Collection, <http://nlp.fi.muni.cz/projekty/eudml/MREC/index.html>

Sojka, P. (ed.): Towards a Digital Mathematics Library. Masaryk University, Paris, France (Jul 2010), <http://www.fi.muni.cz/ sojka/dml-2010-program.html>

Sojka, P., Líška, M.: Indexing and Searching Mathematics in Digital Libraries – Architecture, Design and Scalability Issues. In: Davenport, J.H., Farmer, W., Urban, J., Rabe, F., (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), <http://dx.doi.org/10.1007/978-3-642-22673-1_16>

Sojka, P., Líška, M.: The Art of Mathematics Retrieval. In: Tompa, F., Hardy, M. (eds.) Proceedings of DocEng 2011 Conference. pp. 57–60. ACM. Mountain View, September 2011.

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>

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>

Sylwestrzak, W., Borbinha, J., Bouche, T., Nowiński, A., Sojka, P.: EuDML—Towards the European Digital Mathematics Library. In: Sojka [4], pp. 11–24, <http://dml.cz/dmlcz/702569>

Martin Líška, Petr Sojka, Michal Růžička, and Petr Mravec.
Web Interface and Collection for Mathematical Retrieval.
In Petr Sojka and Thierry Bouche, editors, Proceedings of DML 2011, pages 77–84, Bertinoro, Italy, July 2011. Masaryk University. <http://dml.cz/dmlcz/702604>.